Study Note 2004-03

Understanding and Improving the Assessment of Individual Motivation (AIM) in the Army's GED Plus Program

Deirdre J. Knapp (Editor)
Human Resources Research Organization

Eric D. Heggestad (Editor)
Colorado State University

Mark C. Young (Editor)
U.S. Army Research Institute

United States Army Research Institute for the Behavioral and Social Sciences

January 2004

Approved for public release; distribution is unlimited.

U.S. Army Research Institute for the Behavioral and Social Sciences

A Directorate of the U.S. Army Human Resources Command

ZITA M. SIMUTIS
Director

Research accomplished under contract
for the Department of the Army

Human Resources Research Organization

Technical Review by

William L. Farmer, U.S. Navy Personnel Command, NPRST

NOTICES

DISTRIBUTION: Primary distribution of this Study Note has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, Attn: DAPE-ARI-PO, 5001 Eisenhower Ave., Alexandria, VA 22304-4841.

FINAL DISPOSITION: This Study Note may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this Study Note are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy) January 2004	2. REPORT TYPE Final	3. DATES COVERED (from. . . to) September 30, 2000 – December 31, 2001

4. TITLE AND SUBTITLE Understanding and Improving the Assessment of Individual Motivation (AIM) in the Army's GED Plus Program	5a. CONTRACT OR GRANT NUMBER DASW01-98-D-0047/0025
	5b. PROGRAM ELEMENT NUMBER 310645/331711/331217
6. AUTHOR(S) Deirdre J. Knapp (Human Resources Research Organization), Eric D. Heggestad (Colorado State University), & Mark C. Young (U.S. Army Research Institute) (Editors)	5c. PROJECT NUMBER D730
	5d. TASK NUMBER 266
	5e. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Human Resources Research Organization 66 Canal Center Plaza, Ste 400 Alexandria, VA 22314	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U. S. Army Research Institute for the Behavioral & Social Sciences 5001 Eisenhower Avenue Alexandria, VA 22304-4841	10. MONITOR ACRONYM ARI
	11. MONITOR REPORT NUMBER Study Note 2004-03

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

Mark Young, Contracting Officer's Representative

14. ABSTRACT (Maximum 200 words):

The Assessment of Individual Motivation (AIM) test was developed by the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) to assess work-related temperament characteristics. In February 2000, the Army implemented AIM as a new market-expansion enlistment screening tool under the "GED Plus" program. Under this program, non-high school diploma graduates who might otherwise be ineligible for military service can enlist if they score sufficiently high on the AIM and meet other program requirements. This project addressed several operational issues pertaining to AIM's ongoing use in the GED Plus program. Post-implementation investigations have included (a) a preliminary examination of the operational AIM's validity against attrition under the GED Plus program, (b) the scaling of AIM alternate forms, (c) an examination of variables that might be used to supplement AIMin the prediction of first-term attrition, (d) fairness analyses, and (e) efforts to develop improved ways to score the AIM.

15. SUBJECT TERMS

Selection, Attrition, Turnover, Adaptability Screening

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON (Name and Telephone Number)
16. REPORT Unclassified	17. ABSTRACT Unclassified	18. THIS PAGE Unclassified	Unlimited		Dr. Mark Young (703) 617-0334

FOREWORD

In February 2000, the Army implemented a U.S. Army Research Institute (ARI) test—the Assessment of Individual Motivation (AIM)—as a pre-enlistment attrition screen under the new market-expansion pilot program, GED Plus. The initiation of GED Plus marked the first time AIM had ever been used operationally on a large scale. This report documents preliminary findings of a study designed to evaluate AIM's performance under GED Plus and enhance its performance as an attrition screen for enlisted applicants. The study reported here has been sponsored by the Enlisted Accessions Division, within the Army G-1 (formerly the Office of the Deputy Chief of Staff for Personnel). The findings from this effort have been presented to the Commander, U.S. Army Accessions Command, the Chief, Enlisted Accessions Division, the U.S. Army Accessions Research Consortium, the U.S. Army Accessions Command Attrition Working Group, and the Manpower Accession Policy Working Group.

ARI's Selection and Assignment Research Unit conducts research, studies, and analyses of individual difference measures (of aptitudes, motivations, and other attributes) related to Soldiers' job performance. The primary goal is to improve the Army's selection and classification, promotion, and reassignment of enlisted Soldiers and officers. This document reports on one recent effort to improve enlistment screening procedures for non-high school graduates using a new measure of motivational attributes.

MICHAEL G. RUMSEY
Acting Technical Director

This document contains blank pages that were not filmed.

UNDERSTANDING AND IMPROVING THE ASSESSMENT OF INDIVIDUAL MOTIVATION (AIM) IN THE ARMY'S GED PLUS PROGRAM

EXECUTIVE SUMMARY

Requirement

The Assessment of Individual Motivation (AIM) is a test developed by the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) to assess important work-related temperament characteristics. In February 2000, the Army implemented AIM as a new market-expansion screening tool under the "GED (General Education Development) Plus" program. Under this program, non-high school diploma graduates who might otherwise be ineligible for military service can enlist if they score sufficiently high on the AIM and meet other program requirements. The GED Plus program provides market expansion opportunity in a difficult recruiting environment and allows more opportunities for youth—especially minority youth—to join the Army. The current project was designed to address several operational issues pertaining to AIM's use in the GED Plus program.

Procedure

Under this project, the Human Resources Research Organization (HumRRO) conducted five tasks. Specifically, HumRRO (a) updated research and operational databases needed for AIM's assessment and refinement, (b) computed preliminary estimates of AIM's validity as an attrition predictor using operational data, (c) evaluated and scaled the alternate AIM forms, (d) examined the effects of recruit characteristics on first-term attrition, and (e) managed the coordination and integration of multiple AIM evaluation/refinement efforts conducted by independent consultants working directly for ARI. The consultants (Dr. Fritz Drasgow and Dr. Michael Levine of the University of Illinois, Dr. Paul Sackett of the University of Minnesota, and Dr. Eric Heggestad of Colorado State University) conducted studies of the data from the AIM research databases and served as technical advisors to the project work. Drs. Drasgow and Levine led efforts to explore alternative ways of scoring the AIM and Dr. Sackett took the lead on efforts to investigate test fairness issues.

Findings

The preliminary examination of AIM's operational performance - based on the original scoring procedure - showed that its validity as a predictor of attrition under the GED Plus program was much lower than it had demonstrated in previous work in a research setting. Fairness analyses showed some differential prediction by race/ethnic group, but the investigators hypothesized that there may be other variables that explain this finding. No evidence was found to suggest that AIM's use as an operational screening measure would result in adverse impact for females or minorities. Potential strategies for improving the prediction of first-term attrition, including supplemental predictor variables, reconfiguring the current composite score used for selection, and adopting different scoring strategies (e.g., based on item response theory methods) showed some promise, but most require additional work (conceptualization, data collection, and data analysis) to evaluate more fully. Technical issues associated with the partial ipsativity of the AIM (e.g., how to construct, scale, and evaluate new forms) were identified and discussed.

Utilization of Findings

In addition to documenting how the AIM is working in an operational environment and raising possibilities for alternative scoring approaches, this report is intended to help ARI determine how best to proceed with its future efforts in support of the AIM testing program.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

List of Appendices

List of Tables

List of Figures

PREFACE

Attrition continues to be an expensive and persistent problem for the Army and the other military services. The U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) has been investigating enlisted attrition for many years, for the purpose of helping the Army to better predict and control this phenomenon. One primary objective has been to identify and measure the characteristics and motivational attributes of youth who are most likely to succeed as Soldiers. This has been a challenging task for two reasons. First, the phenomenon of attrition itself has been rather elusive. It represents a complex outcome affected by the interaction of many different factors (individual, organizational, and extra-organizational). In addition, attempts to capture individual differences relevant to attrition often rely on self-report measures because it is difficult to capture this information in other ways. Such self-report measures are susceptible to response distortion (faking) which degrades the measurement process. These are some of the challenges our team of researchers faced in the investigations reported here.

This is the first public report documenting ARI's program to support the pre-enlistment attrition screen being used under the recruiting market expansion pilot program, GED Plus. The primary purpose of our program has been to assess how well the Assessment of Individual Motivation (AIM), a self-report measure, performs in this operational context and to explore ways of refining or augmenting AIM to enhance its utility as an attrition screen for enlisted applicants. This study has been sponsored by the Enlisted Accessions Division, within the Army G-1 (formerly the Office of the Deputy Chief of Staff for Personnel). Most of this effort was carried out through contracts with the Human Resources Research Organization (HumRRO) and several university faculty consultants hired through the Consortium of Universities Research Fellows Program. I served as the ARI contract monitor.

While this report focuses on AIM's use as a pre-enlistment attrition screen, research findings on AIM in other (post-enlistment) applications have also been encouraging. These include maturity screening for correctional specialists, the selection of Army recruiters, and predicting success among Explosive Ordnance Disposal (EOD) trainees. See White and Young (2001) for a short summary of these efforts.

Past failures along with related concerns about faking and coaching have made it difficult to implement self-report measures of motivational attributes for the selection of military personnel. However, the Army's implementation of the experimental GED Plus program in February 2000 marked the beginning of AIM's operational use for pre-enlistment screening. Although a previous large-scale validation effort (1998-1999) showed that AIM was a reliable predictor of first-term attrition, Soldiers in that effort had completed AIM for research purposes only. At the time of testing, those new recruits were specifically told that their AIM scores would have no effect on their Army careers. Therefore, prior to AIM's use in the GED Plus Program, no one knew how well the earlier validation results from the research setting would generalize to the more challenging "operational" setting. Unlike the new recruits in our earlier validation research, applicants tested under GED Plus would have reason to be highly motivated to perform well on AIM. These applicants would be required to meet a certain AIM cut score in order to enter the Army or be eligible for special incentives that might be worth thousands of dollars. The work and findings presented in this report highlight the many difficulties and challenges of implementing a quasi-ipsative, self-report measure (like AIM) for the purpose of attrition screening. The concern for maintaining validity in an operational environment is just one of many complex issues addressed here.

The report's findings on AIM's performance under GED Plus are preliminary. This is because the program had only been in place for a short time, and much more operational data have become available for analysis since the timeframe of the contract effort. Importantly, as our program has evolved quickly over a short time, there have been a number of new developments and findings which are outside the timeframe of the work documented here. For this reason, ARI has inserted editorial comments at the beginning of each chapter to provide a broader context and help bridge the gap between the documented efforts and the current state of the project.

AIM is known to be resistant, but not impervious, to faking. For this reason, we have disguised the names of AIM scales (identified by a letter) in those situations where revealing their identities could potentially contribute to test compromise (Chapters 3, 4, 5, 7, & 9). We have taken a similar approach to disguising the content of survey items (Chapter 5) which might be considered for operational use sometime in the future. A restricted version of this report which reveals this sensitive content (Knapp, Waters, & Heggestad, 2002; ARI Study Note 2002-02; July 2002) is available on a limited basis.

We appreciate the encouragement and support of the many individuals and organizations that have contributed to this effort. They include HumRRO, RAND, and the Air Force Office of Testing and Survey Policy. As HumRRO's project director, Deirdre Knapp had a lead role in coordinating the various research efforts and integrating them into a meaningful report. This was a daunting task. Key HumRRO project staff included Jeff Barnes, Rod McCloy, and Dan Putka. HumRRO also assisted in facilitating an independent external review (by consultants) of our AIM program (January – March 2002). We are also grateful to the personnel at the Air Force Office of Testing and Survey Policy and Bruce Orvis (RAND) for generously sharing data and time with ARI in support of our program. ARI's university faculty consultants also contributed greatly, both through their research efforts reported here, and their role as technical advisors in the pre-implementation research. These consultants include Fritz Drasgow, Michael Levine, Bruce Williams, Paul Sackett, and Eric Heggestad.

Finally, we want to comment on the utilization of the findings documented in this report. The findings from this effort contributed to our development of a new attrition screen (AIM with revised scoring plus supplemental measures) which ARI proposed as a replacement for the existing operational AIM score under the GED Plus Program. ARI made this recommendation in a briefing to LTG Dennis Cavin, Commander, U.S. Army Accessions Command in July 2002. As the pilot program progresses through fiscal year 2003, the Army will decide the future of the GED Plus Program and what (if any) operational screen will be used as part of the program.

Mark C. Young, ARI
August 2003

CHAPTER 1. INTRODUCTION: AN OVERVIEW OF AIM AND THE PRELIMINARY EFFORTS TO SUPPORT ITS OPERATIONAL USE

Mark C. Young
U.S. Army Research Institute for the Behavioral and Social Sciences (ARI)

Rodney A. McCloy
Human Resources Research Organization (HumRRO)

Brian K. Waters
Human Resources Research Organization (HumRRO)

Leonard A. White
U.S. Army Research Institute for the Behavioral and Social Sciences (ARI)

This chapter summarizes how AIM evolved from its initial development (1992 – 1996) as a self-report measure of motivational attributes, to eventually become an operational attrition screen (2000) under the Army's experimental pilot program, GED Plus. Previous data obtained in a research context showed AIM to be a valid predictor of attrition and resistant to faking. However, the transition to operational use provides the most critical test of AIM's performance, since AIM had not previously been used in a large operational context. The chapter ends with an overview of the remaining chapters. The efforts reflected in these chapters pertain to (a) determining the limits of AIM's performance in the operational setting and (b) exploring ways to address these limitations in order to reduce attrition through personnel selection.

Background: The Assessment of Individual Motivation (AIM)

Over the past decade, there has been a resurgence of interest in using personality (or temperament) assessments for personnel decisions. The military has been no exception. Many studies in military settings have shown that motivational attributes are important predictors of both Soldier attrition and motivational aspects of performance. For example, during the early 1980s, the Army developed the Assessment of Background and Life Experiences (ABLE), a test designed to assess seven personality constructs (Dependability, Adjustment, Internal Control, Work Orientation, Agreeableness, Leadership, Physical Conditioning) that were expected to be useful for predicting job success. Personality and motivation measures – like ABLE – have not been used in the Army's selection program for enlisted personnel. While the ABLE had predictive validity for certain aspects of job performance and for first-term attrition (Campbell & Knapp, 2001; White, Young, & Rumsey, 2001), research indicated that the test was also susceptible to the effects of faking and coaching (see Hough, Eaton, Dunette, Kamp, & McCloy, 1990; White et al., 2001; Young, White, & Oppler, 1991). Studies designed to assess the effect of faking or coaching demonstrated that the predictive validity of the ABLE was near zero for job-related criteria when test-takers were coached or instructed to fake their scores (White et al., 2001; Young, White, & Oppler, 1992). In part on the basis of these results, the ABLE was never operationally implemented in the Army.

The U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) developed the Assessment of Individual Motivation (AIM) in 1996 in an effort to build an assessment tool that would provide scores for the same constructs measured by ABLE, but in a way that was resistant to the effects of faking and coaching.

Development of the AIM

Over a 4-year period, seven developmental versions of AIM were administered to a total of approximately 5,000 new Army recruits. Over several iterations, test forms were administered and refined until the prototype AIM form was finalized and evaluated in 1996.

Test Construction

The strategy for developing AIM differs from that of ABLE in several significant ways (White & Young, 1998, 2001; Young & White, 1998). First, AIM uses a forced-choice format to reduce item transparency and place constraints on faking. AIM items consist of four statements (a tetrad) that may describe an examinee's past behavior in familiar situations. Two of these statements are worded positively (often indicating a high standing on the construct) and two are worded negatively (often indicating a low standing on the construct). For each item, respondents are asked to select the one statement (stem) which is *most like* them, and the one statement which is *least like* them. A quasi-ipsative scoring method is used to generate four construct scores for each item (i.e., one score for each stem). Scale scores are obtained by summing—across items—the scores for stems measuring the same construct.

Another important strategy in AIM's development was to create items which focused as much as possible on behaviors, thereby making them like biodata items. This contrasts with ABLE, which contains biodata items, but also items relating to personal attitudes, affect, and traits. Research from ABLE was very useful in identifying past experiences and behaviors linked to the target constructs, and therefore helped to guide ARI's development and revision of the AIM items.

In constructing AIM, ARI also sought to minimize the effects of social desirability on item responding. This was done by attempting to equate stems within the same item on social desirability, and revising (or replacing) stems that had very high endorsement rates or were highly correlated with an index of response distortion. ARI used traditional item statistics (e.g., item-total correlations and coefficient alpha reliability estimates) as a primary means for evaluating item and scale-level performance of AIM throughout its development and refinement.

Description of AIM

The final AIM form consists of 27 forced-choice items which provide scores on six of the seven ABLE constructs (all but Internal Control): Dependability, Adjustment, Work Orientation, Agreeableness, Leadership, and Physical Conditioning. Definitions for these six constructs, some of which overlap with the Five-Factor Model (see White et al., 2001), are presented in Table 1.1. AIM also contains a Validity scale for assessing the amount of response distortion that may be reflected in the measure. Figure 1.1 shows a sample AIM item, indicating the stems selected as *most like* (M) and *least like* (L) the respondent.

Table 1.2. Definitions of the AIM Scales

Title	Definition
Work Orientation	The tendency to strive for excellence in the completion of work-related tasks. Persons high on this construct seek challenging work activities and set high standards for themselves. They consistently work hard to meet these high standards.
Adjustment	The tendency to have a uniformly positive affect. Persons high on this construct maintain a positive outlook on life, are free of excessive fears and worries, and have a feeling of self-control. They maintain their positive affect and self-control even when faced with stressful circumstances.
Agreeableness	The tendency to interact with others in a pleasant manner. Persons high on this construct get along and work well with others. They show kindness, while avoiding arguments and negative emotional outbursts directed at others.
Dependability	The tendency to respect and obey rules, regulations, and authority figures. Persons high on this construct are more likely to stay out of trouble in the workplace and avoid getting into difficulties with law enforcement officials.
Leadership	The tendency to seek out and enjoy being in leadership positions. Persons high on this scale are confident of their abilities and gravitate towards leadership roles in groups. They feel comfortable directing the activities of other people and are looked to for direction when group decisions have to be made.
Physical Conditioning	The tendency to seek out and participate in physically demanding activities. Persons high on this construct routinely participate in vigorous sports or exercise, and enjoy hard physical work.

____	(A) I have almost always completed projects on time.
____	(B) I have not exercised regularly.
M	(C) I have enjoyed coordinating the activities of others.
L	(D) I have a hard time feeling relaxed before an important test.

Figure 1.1. Sample AIM item.

Preliminary Research Findings

Preliminary research indicated that AIM did in fact predict first-term attrition, while being less susceptible to the effects of faking and coaching than the ABLE (results of this research have been summarized by White & Young, 1998; White et al., 2001; Young & White, 1998). Recruits' AIM scores were also shown to be positively correlated with self-report measures of their adjustment to the military, confidence in their ability to complete Basic Training, and commitment to complete their enlistment term. In addition, AIM scores were shown to be similar across gender and racial groups. Those positive preliminary findings increased the Army's interest in using the AIM as a pre-enlistment screen for attrition.

In July 1997, ARI convened an external panel of testing experts to review these preliminary findings and evaluate AIM's potential as an Army pre-enlistment screening measure. The panel concluded that although the findings were promising, much more data would be needed—including data from a large-scale predictive validation study—before making a recommendation on implementation. As a result, ARI initiated a program of research to "(a) determine whether the use of AIM for pre-enlistment screening would be viable for the Army, and (b) remove obstacles to its implementation" (Young & Rumsey, 1998, p. 1).

Pre-Implementation Research Program

Under the sponsorship of the Army's Deputy Chief of Staff for Personnel (now Army G-1), ARI began its AIM Pre-Implementation Research Program (1998–1999) to assess whether AIM would be viable for screening enlisted Army applicants. To consider implementing AIM in an operational environment, the initial promising results needed replication in a much larger and more representative sample of Army accessions. At the same time, any decision to implement a new selection screen requires estimates of validity, utility, and adverse impact. Finally, operational use of any selection measure requires that there be more than one form of the measure available. Alternate AIM forms needed to be constructed, and their performance compared against the original form.

In the pre-implementation research program, researchers (a) validated AIM against initial entry training attrition, (b) developed and evaluated alternate forms, (c) evaluated AIM's potential adverse impact when used for enlistment screening, and (d) assessed AIM's resistance to faking.

Technical Advisory Panel

ARI assembled a panel of five testing experts to provide periodic external reviews and technical guidance during the AIM pre-implementation research effort. Several of these panel members subsequently took more active roles in the post-implementation research. The panel first met with ARI and HumRRO in November 1998. At mid-course, in February 1999, the panel met again to review the program's preliminary findings and make recommendations on its future course. Individual panel members provided input on technical issues on an informal basis throughout the duration of the program. Well before the conclusion of the pre-implementation testing period, the panel recommended that the Army proceed with an Initial Operational Test and Evaluation (IOT&E) of AIM.

Item Development and Data Collection

Researchers developed items for two alternative forms of the original AIM (see Figure 1.2 for a flow chart of the activities discussed in the next several sections). The process of creating the alternate forms began by researchers developing pools of stems for each of the six AIM constructs and for the Validity scale. In addition, combinations of the scales were analyzed for possible use as composites for predicting first-term attrition. One such measure, the "Adaptability" composite, comprises three of the content-based scales (Heggestad, Young, Strickland, & Rumsey, 1999).

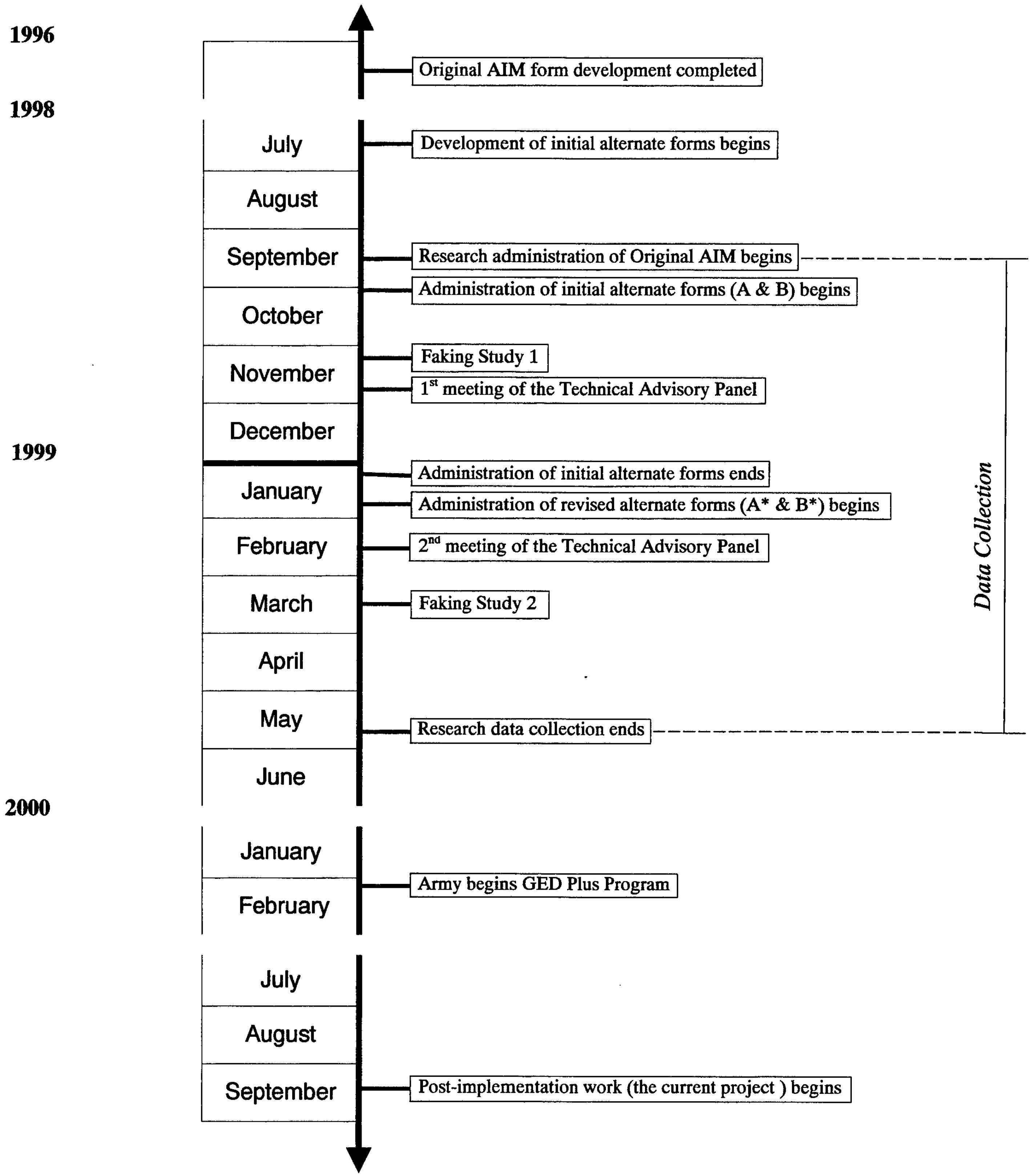


Figure 1.2. AIM background flowchart.

AIM was administered to over 22,000 recruits entering the Army from September 1998 through May 1999. Recruits took the AIM as part of their in-processing activities at all six Army Reception Battalions and were told that their test scores would be used only for research, and would not have any impact on their Army careers.[1] Tested recruits were later tracked to determine their attrition status at 3, 6, and 9 months. (A detailed description of the resulting Army AIM Research Database is provided in Chapter 2 of this report). Researchers then evaluated the AIM's ability to predict attrition for each time period and the ability of the alternate forms to duplicate the results from the original AIM form.

The first (developmental) versions of the alternative forms (A & B), the "initial alternate forms," were administered from October 1998 through early January 1999. After evaluating their performance relative to the original AIM, Forms A and B were revised to create the "revised alternative forms" (A* & B*) and then administered from late January 1999 through May 1999. Note that the tested recruits always took the original AIM before completing either of the two sets of alternate forms within a single testing session. No retest data were collected.

Research Findings

The results of analyses designed to evaluate AIM's performance in the pre-implementation research sample were presented in Heggestad et al. (1999) and Young, Heggestad, Rumsey, and White (2000). Three attrition samples were defined and six alternative attrition composites evaluated. Cases were coded "1" if the Soldier attrited and "0" if the Soldier did not attrit.

Validity of AIM Against Attrition

In general, the potential AIM composite scores were found to have modest validity against attrition in each of the three samples. The attrit group had mean scores that were generally one-half of a standard deviation below those of the non-attrit group. The point-biserial correlations between the composites and attrition status fell between -.10 and -.20. The negative sign of the validity coefficients indicates that low AIM scores are associated with higher attrition rates.

The decile plot of AIM Adaptability Composite scores against 3-month attrition for Tier 1 (High School Diploma Graduate) recruits (n = 14, 500; r = -.12) is presented in Figure 1.3 (from Young et. al., 2000). The deciles (shown along the x axis) depict 10 ranked AIM score categories – with approximately 10% of the cases falling within each decile. Recruits scoring below the 11th percentile on AIM were assigned to decile 1(the lowest ranking), whereas those scoring in the top 10% of the AIM score distribution were assigned to decile 10 (the highest ranking). The vertical bars represent the attrition rate observed within each AIM decile rank.

[1] It was important to test at *all* Reception Battalions because military specialties and trainee characteristics (e.g., gender) vary by location.

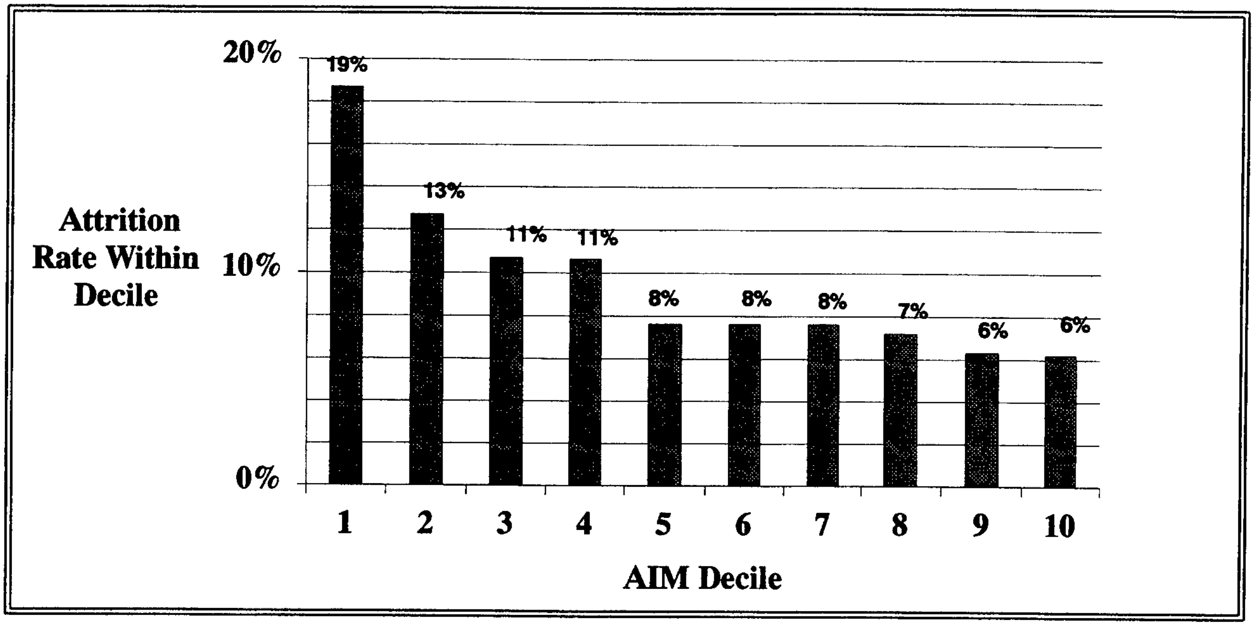


Figure 1.3. Relationship between AIM and 3-month attrition among Tier 1 Regular Army accessions (n = 14,500).

Figure 1.3 shows AIM's nonlinear relationship with attrition, which is consistent with findings from Tier 2 (Alternative Education Credential) data (not presented here). While AIM scores falling in the lowest decile are especially indicative of attrition risk, attrition levels off at a relatively low rate around the 50th percentile. The decile plots of AIM against 6-month (r = -.14) and 9-month (r = -.15) attrition were highly similar to that shown in Figure 1.3, except that the bars (representing attrition rates within deciles) were taller – indicating the higher base rates of attrition associated with these more mature criterion measures. Highly similar findings were also obtained from a U.S. Air Force sample of 15,000 airman trainees who completed AIM in 1997 and were tracked over time to determine their attrition status. (A sample taken from the Air Force AIM Research Database described in Chapter 2 of this report). Not surprisingly, this same type of nonlinear relationship with attrition was also noted for AIM's precursor, ABLE (White et al., 2001).

Also like ABLE, AIM was shown to have incremental validity beyond the measures currently used for enlistment screening (e.g. Educational Tier and Armed Forces Qualification Test [AFQT] scores). In addition, the magnitude of AIM's relationship with 3-month attrition was shown to compare favorably with that of high school diploma status. This was an encouraging finding given that high school diploma status is generally considered to be the single best predictor of military attrition. (This positive finding continues to hold up, even as ARI has tracked attrition through 18 months).

Although the magnitude of AIM's correlation with attrition is modest, it is comparable to that of its precursor, ABLE. Past utility analyses with ABLE have shown that these low levels of predictive validity can have positive utility within those market segments (such as nongraduate youth in educational Tiers 2 and 3) where recruiting costs are relatively low (White, Nord, Mael, & Young, 1993).

Validity of AIM Against Training Performance

The relationship between AIM and training performance was examined for subsamples of Soldiers tested at two Reception Battalions (Heggestad, Young, Strickland, & Rumsey, 1999). Trainees identified as having one or more Articles 15—an indicator of indiscipline—had lower scores on the AIM Dependability scale (effect size = -.25) than those without such infractions. The AIM scores of new recruits participating in a special program for those with motivational and disciplinary problems were also examined. Compared with nonparticipants, these trainees were found to have lower AIM scores, especially on the Dependability, Work Orientation, and Agreeableness scales (with effect sizes of -.27, -.22, and -.24, respectively). Research findings also showed no evidence that the use of AIM as an attrition screen would result in adverse impact for females or minorities.

Evaluation of Alternate Forms

Heggestad et al. (1999) compared the initial alternate forms to the original AIM at the stem and scale levels. They also evaluated the validity and utility of the initial alternate forms relative to the original form. Although some differences between the initial alternate forms and the original AIM form were noted, overall the various forms were highly similar to each other. Across forms, there was a strong correspondence between scales measuring the same construct[2], and the AIM composites had very similar validities against attrition. In addition, analyses evaluating the utility of selection decisions using two hypothetical cut scores showed similar results across forms.

Although the revisions made to the initial alternate forms were intended to improve their correspondence to the original form, there remained evidence of somewhat different response tendencies. However, an examination of the revised alternate forms at the scale level indicated that the means, scale intercorrelations, and internal consistency reliability coefficients were generally comparable to those observed for the original AIM form. An evaluation of the same-scale correlations across the forms indicated good construct correspondence between the scales.

A limited assessment of validity and utility for the revised alternate forms was made against a 3-month attrition criterion. A comparison of the utility of the composites indicated that the quality of decisions made at two cut scores was slightly better for the original AIM.

Effects of Faking on AIM

Finally, Heggestad et al. (1999) assessed the susceptibility of the original AIM and the alternate forms to the effects of faking through two studies—one involving the original alternate forms and one involving the revised alternate forms. Both studies evaluated the effect of two faking conditions: (a) a *fake maximal instruction condition*, intended to assess the degree to which AIM could be faked; and (b) a *fake operational condition*, intended to assess the amount of faking that might be expected in an operational environment. The two conditions were identical except for one important difference: The fake operational condition included a warning statement about the detection of and consequences for faking, whereas the fake maximal condition did not. The test administrator read the following instructions aloud to the subjects in both conditions:

---

[2] Same-scale correlations across forms, corrected for attenuation, ranged from .77 to .99, with only 3 falling below .86; 11 were greater than .90.

The Army is developing a questionnaire that may eventually be used to help determine who is let into the Army. We are going to have you complete the questionnaire today, but we want you to do it in a little different way.

We want you to pretend that you have to take this questionnaire to see if you are good Army material. Imagine also that the recruiter has told you that if you get a good score you will be let into the Army.

At this time, please open your test booklets to the instructions on the inside front cover. I will now read the standard instructions for the questionnaire. In completing the questions, remember that you are to pretend that you want to get into the Army and that you need to get a good score on this questionnaire to do so.

Subjects in the fake operational (but not the fake maximal) condition were given the same warning as those given the standard instructions. This warning, which appeared in the test booklets and was read aloud to the subjects, was worded as follows:

Your answers make it possible to detect if you are describing yourself as better than you really are. Dishonest responding will lower your score, so it is important to answer truthfully.

No type of warning against faking was given to those in the fake maximal condition.

Effect sizes of AIM scale scores were computed for both faking conditions (compared against the standard instructions; n = 1,684), and separate analyses were conducted for the subsamples of recruits who completed the initial versus the revised alternate forms. Recruits in each subsample completed both the original AIM and one set of the (initial or revised) alternate forms. Table 1.2 presents the findings for the recruits who completed the initial alternate forms. To summarize the data more succinctly, the content scale effect sizes shown in the table represent the mean effect sizes across the six content scales.

Table 1.2. Effect Sizes By Faking Condition in Initial Alternate Form Sample

Test Form	Instructional Condition	
	Fake Operational (n = 648)	Fake Maximal (n = 538)
Original AIM		
Content Scales	0.15	0.33
Validity Scale	0.30	0.93
Form A		
Content Scales	0.19	0.23
Validity Scale	0.17	0.67
Form B		
Content Scales	0.14	0.14
Validity Scales	0.24	0.71

Table 1.2 shows that, although modest score inflation was detected, it was less than has been found for other non-cognitive assessment instruments. Overall, the magnitude of score inflation found for the initial alternate forms was similar to that of the original AIM. The data also shows a pattern of higher score escalation for the Validity scale (which was developed to detect response distortion) than for the content scales. The faking instructions were found to have a minimal impact on the correlational structure of the scales of both the original AIM and the alternate forms.

In a follow-up project, Young, White, and Heggestad (2001) tracked the subjects from these faking experiments (n = 1,597) to assess the impact of faking on the validity of the original AIM's Adaptability Composite against 18-month attrition. The effect sizes for the Adaptability Composite in the fake operational (n = 899) and fake maximal conditions (n = 698) were .09 and .39 respectively. Consistent with past research (including the findings shown in Table 1.2), warning statements were effective in suppressing score inflation among respondents who were encouraged to fake. Moreover, whether or not subjects were given warnings about the detection of faking, instructions encouraging them to inflate their scores had no statistically significant impact on attenuating AIM's validity as a predictor of 18-month attrition, relative to the standard (honest) instructions (n = 22,593). The authors of this research noted the uncertainty as to how well the results from this "simulated applicant" sample from a research setting would generalize to an operational setting where applicants would be highly motivated to score well.

Implementation of AIM Under the GED Plus Program

When the Army first began exploring the feasibility of implementing AIM (1997), AIM was primarily viewed as a potential screening tool for reducing enlisted attrition. However, soon after the pre-implmentation research program began, it became apparent that the growing recruiting difficulties faced by the Army would most likely preclude any application of AIM that reduced the pool of available candidates. This realization led to the discussion of options for AIM's use which focused on pre-enlistment screening *within an expanded recruiting market.* Using AIM in the context of an expanded recruiting market would help to offset the added recruiting burden associated with screening. Such an application could make AIM screening viable in those markets where recruiting costs are relatively low.

In February 2000, the Army implemented AIM as a new market-expansion screening tool under the "GED (General Education Development) Plus" program. The GED Plus Program provides market expansion opportunity in a difficult recruiting environment and allows more opportunities for youth to join the Army. This is particularly the case for minorities, some of whom may have dropped out of school for cultural or financial reasons (e.g., to help support the family). Some educational Tier 2 and Tier 3 (No Education Credential) recruits make excellent Soldiers and one goal is to use AIM to help identify applicants who are most likely to succeed. Historically, first-term attrition rates have been lower in Tier 1 than in Tiers 2 and 3. There was – and continues to be – little difference between Tier 2 and Tier 3 attrition rates. Most of the Tier 1 applicants for the Army went to 4-year traditional high schools; home schoolers also have been placed in this category on an experimental basis.

The success of the GED Plus pilot program is being evaluated as part of a large Army experiment. The program is available to Tier 2 applicants at all recruiting locations, but Tier 3 applicants are only selected at 40% of locations. Program candidates are being tested on AIM at Military Entrance Processing Stations (MEPS) nationwide, on an ongoing basis. At the end of the 3-year pilot test, the Army will decide whether or not to adopt the program on a permanent basis. The RAND Corporation is conducting an independent evaluation of the entire GED Plus Program. RAND is interested in market expansion, but like ARI, is also looking at factors related to selection. Initial findings of the GED Plus Program indicate that it is attracting minorities and has been useful in helping the Army to meet its mission of new enlisted accessions.

The Department of Defense caps the combined Tier 2 and Tier 3 accessions at 10%, a limit the Army adheres to. The GED Plus Program allows for qualified Tier 2 and 3 recruits to enter the Army without counting against the 10% cap on nongraduates. While in the Delayed Entry Program (DEP), GED Plus qualified Tier 3 recruits complete adult education programs that lead to a GED high school equivalency diploma. These recruits are allowed to enter active duty once they have completed the GED.

To enter the GED Plus Program, Tier 2 and 3 applicants must have scored 50 or higher on AFQT, score well on the AIM Adaptability Composite (approximately upper 75%—based on an earlier research sample of Army recruits), have no moral waiver (except minor offenses) and no drug or alcohol waiver. They must have disenrolled from high school by their own choice and their ineligibility to return to high school must be due to age.

The AIM testing instructions for GED Plus applicants include a warning statement about the detection and consequences for faking. This warning statement is identical to the warning statement used in the data collection for the AIM Pre-Implementation Research Program.

The Current Project

The current project was designed to address several operational issues pertaining to AIM's ongoing use. Specifically, the objective of this effort was to evaluate and refine AIM and support its continued operational use under the GED Plus program. Under this project, HumRRO (a) updated research and operational databases needed for AIM's assessment and refinement, (b) estimated AIM's validity as an attrition predictor using operational data, (c) evaluated and scaled the alternate AIM forms, (d) examined the effects of recruit characteristics on first-term attrition, and (e) managed the coordination and integration of multiple AIM evaluation/refinement efforts being conducted both under this contract and by independent consultants working directly for ARI. The consultants (Dr. Fritz Drasgow and Dr. Michael Levine of the University of Illinois, Dr. Paul Sackett of the University of Minnesota, and Dr. Eric Heggestad of Colorado State University) conducted studies of the data from the AIM research databases and served as technical advisors to the project work.

Purpose and Overview of Report

This report documents investigations conducted to support AIM's continued operational use. It covers work performed by ARI consultants and HumRRO investigators from September, 2000 through July, 2001.

The report is divided into nine separately-authored chapters, plus consolidated references and appendices across the entire report. This chapter, *Introduction,* describes the background research on AIM.

Chapter 2, *Overview of AIM Databases,* describes three AIM databases: the Army Grand Research Database, the Air Force AIM Research Database, and the Army Operational Database. These databases were those available to the investigators at the time of work reported here. Subsequent updates have since been made.

Chapter 3, *Preliminary AIM Validation Based on GED Plus Program Data*, describes an investigation of the validity of the AIM for predicting Soldiers' 3-, 6-, and 9-month attrition status using operational AIM data gathered as part of the GED Plus program. A secondary purpose was to examine the potential utility that implementing different cut scores might have in terms of decision quality (e.g., hit rates, false acceptance rates).

Chapter 4, *AIM Alternate Form Selection and Scaling,* describes comparative research database analyses of the original and alternate (initial and revised) AIM forms. It also reports the results of scaling the alternate forms to the original AIM.

Chapter 5, *Effects of Recruit Characteristics on First-Term Attrition,* reports comprehensive analyses of attrition in the research databases. The analyses examined the efficacy of the AIM, in combination with other variables (e.g., Armed Services Vocational Aptitude Battery [ASVAB] subtest scores, survey items from ARI's First Term Project), for identifying Soldiers who are high risks for attrition during their first term of enlistment. The specific focus was on predicting 18-month attrition status among two groups that had been identified by past research as being at particularly high risk for attrition, namely, Soldiers classified in educational Tier 2, and female Soldiers classified in educational Tier 1 (Sipes, Strickland, Laurence, DiFazio, & Wetzel, 2000). The primary predictor variables of interest in this study can be classified into three groups, namely: (a) the AIM scales and Adaptability Composite; (b) operational predictors (specifically, ASVAB scores, age, years of education, body mass index, entry pay grade); and (c) predictors from the Soldier Reception Survey that could potentially be used operationally, or for purposes of identifying Soldiers in need of subsequent attrition-reducing interventions. The project investigated the efficacy of these three groups of variables for predicting Soldiers' 18-month attrition status.

Chapter 6, *AIM Adverse Impact, Differential Validity, and Differential Prediction,* addresses issues in the broad domain of AIM "fairness." Sackett and Laczo examine three issues: (a) subgroup differences and adverse impact, (b) differential validity of AIM by gender and race in predicting attrition, and (c) differential prediction (or predictive bias) by gender and race in predicting attrition.

Chapter 7, *Alternative Methodologies for Predicting Attrition in the Army: The New AIM Scales*, examines alternative methods to predict attrition using the AIM data. The analyses include (a) basic analyses (item-total correlations, coefficient alpha at the scale level, principal components analysis, and correlations between each stem and each scale with attrition), (b) multiple linear and logistic regression of retention onto the AIM scales, (c) classification tree modeling of attrition with the AIM scales at 12 months and, (d) item response theory (IRT) modeling of the polytomously-scored AIM scales and optimal classification via the Neyman-Pearson lemma and the IRT models applied to the scales.

Chapter 8, *Robust Modeling and Optimal Classification for AIM,* presents a model-based approach to AIM usage. Using IRT methods, Levine and Williams determined the power of the most effective way to use AIM item-level responses to predict 12-month attrition. They also set out to develop a robust polytomous model for AIM that could serve as a counter-measure to response distortion. A detailed model that simultaneously models AIM's partially ipsative format and the interaction between all six of AIM's components defends against transparency by supporting appropriateness measurement (i.e., determining which and how many examinees are currently faking) and robustification (estimating the AIM score an individual examinee would have obtained if he or she had answered every item honestly).

Chapter 9, *Summary and Future Directions,* summarizes the technical chapters in the report and discusses future directions for investigations related to the use of AIM.

CHAPTER 2: OVERVIEW OF AIM DATABASES

Jeffrey D. Barnes
Human Resources Research Organization

Eric D. Heggestad
Colorado State University

Mark C. Young and Leonard A. White
U.S. Army Research Institute

When ARI's post-implementation effort began in September 2000, AIM screening under the GED Plus program had been ongoing for 7 months. It took time for us to track program applicants long enough to establish their 3-, 6-, and 9-month attrition status. This was critical for developing an appropriate "operational" database for the initial evaluation of AIM's validity as an attrition screen. Within the timeframe of the investigations reported here (September 2000 – July 2001) we were able to obtain data for only a few thousand cases with early attrition criteria. Therefore, our investigations with these GED Plus data are considered preliminary. Since July 2001, there have been updates and refinements to the operational AIM database that are not reported here.

The more mature Army and Air Force AIM research databases that had previously been developed under our AIM Pre-Implementation Research Program (1998-1999) were also updated and refined by HumRRO under this current effort. We relied heavily on these research databases for most of the investigations presented in this report. This chapter describes all three databases used in the analyses conducted by the various authors.

The analyses to be presented in this report were conducted using data from three primary sources. These are:

- Army AIM Grand Research Database
- Air Force AIM Research Database
- Army Operational Database

The purpose of this chapter is to provide an overview of the content and construction of each database, including basic descriptive information and explanations of key selection and control variables.

Army AIM Grand Research Database

New recruits report to a Reception Battalion for 2-3 days of in-processing prior to the start of basic training. Between September 1998 and May 1999, the AIM was administered to most new recruits at the six Army Reception Battalions. Test responses, identification information, and demographic variables were recorded on custom designed answer forms that could be read using optical mark reader (OMR) scanning equipment. Although OMR scanning is highly accurate (i.e., less than 0.1 % error), the raw data files were processed through a series of edit routines to identify missing or incomplete Social Security Numbers (SSNs), multiple or missing responses to AIM questions, and multiple responses to other variables. When such errors were identified, the answer sheets were retrieved and visually inspected. We recorded corrections on the error report document and key-entered into the raw data files. In total, 37,295 Soldiers completed the AIM during this research phase.

The research sample can be broken down into six test administration conditions involving three versions of the AIM and two instructional conditions. Table 2.1 shows the distribution of data records by test administration condition. The three versions of the AIM included (a) administration of the original AIM alone, (b) administration of the original AIM and two initial alternate forms (A and B), and (c) administration of the original AIM with two revised alternate forms (A* and B*). When the alternate forms (initial and revised versions) were given, the original AIM was always administered first. The order of administration of the alternate forms was balanced within each session. The new recruits each participated in a single test administration session.

Each version of AIM was administered under two instructional conditions: standard instructions and fake good instructions (a detailed explanation of the fake good instruction set is provided in Heggestad, Young, Strickland, & Rumsey, 1999). As shown in Table 2.1, 2,768 individuals completed AIM under these special instruction conditions designed to evaluate the potential for response distortion.

Table 2.1. Sample Sizes for AIM Versions

AIM Version	Standard Instructions	Faking Experiment	Total *n*
Original AIM only	21,097	209	21,306
AIM & Initial Alternate Forms (Forms A and B)	2,706	1,438	4,144
AIM & Revised Alternate Forms (Form A * and B*)	10,724	1,121	11,845
Total *n*	34,527	2,768	37,295

To ensure a consistent data structure across the various testing conditions, all data elements were mapped to a single condition. The standard selected was the 27-item original AIM followed by the 26-item initial alternate Form A and the 26-item initial alternate Form B. It should be noted that in the original AIM, the first item presented is treated as a practice item and, therefore, not scored. No practice items were included in the initial alternate forms, which accounts for there being one item fewer per form on the alternates. Next, item level scoring was performed. During this step, the length of the forms was equalized to 27 items by assuming that the first item of the original AIM was present as the first item on each alternate form. This simplified the computer code required to generate item-level scores, and provided a unified variable naming structure. Scale scores were then calculated and added to the database.

Consistent with procedures outlined by ARI, when less than 10% of the data from a scale were missing, the average of non-missing items in the scale was imputed for the missing values. It is important to note that imputations of item-level scores only took place during the calculation of the various scale scores; the database does not contain any imputed item-level responses. Scale scores were not calculated if more than 10% of items contributing to the scale were missing.

A primary goal of the investigation was to determine whether AIM, in conjunction with other data collected at enlistment, assisted in identifying Soldiers more likely to separate from the service prior to completion of their obligation. The two primary sources of enlistment data were the U.S. Military Entrance Processing Command (USMEPCOM) Integrated Resource System (MIRS) and the Enlisted Master File (EMF). ARI chose a comprehensive list of desired variables from each database that would be merged with the AIM research data to each file based on SSN matching. The Army MIRS extract is updated monthly, while the EMF is updated quarterly. The first match obtained from each file provided the variable values at enlistment. We have subsequently matched these values to quarterly EMF files to obtain separation information – specifically, date, type, and character of separation, as well as the separation program designator. The database update used for the analyses described in this report captures separations through 31 December 2000.

The EMF contains records for Regular Army Soldiers only. Excluding the faking experiment participants, 22,859 (66.3%) of the remaining 34,527 records were Regular Army Soldiers; 7,789 (22.6%) were from the Army National Guard; and 3,848 (11.1%) were from the Army Reserves. Component membership of 31 individuals could not be determined. Of the Regular Army Soldiers, 322 cases could not be matched with the EMF database, and, of these, 134 cases could not be matched to the MIRS database either. Thus, the database includes 22,537 Regular Army Soldiers who took one form of the AIM under standard instruction conditions.

Attrition Criterion Variables

Attrition criteria are constructed as part of each quarterly update. First, we calculate a theoretical maximum number of days of service – this is defined as the time between entry onto active duty and the EMF file cutoff date (e.g., 31 December 2000). Next, we compute the time in service (TIS). For those who have not separated, TIS equals the theoretical maximum. For those who have separated, TIS represents the number of days between enlistment and separation. If the Soldier has separated, the attrition flag is set to the value 1. This attrition flag is reset to a special missing value if the separation was due to death or entry into an officer training program. The flag is reset to 0 if the separation was the result of an immediate reenlistment.

In addition to the overall attrition flag, we established 12 attrition status variables corresponding to lengths of service starting at 3 months (91 days) and increasing in 3-month increments to a maximum of 36 months. To be included in the cohort corresponding to a specific attrition criterion, a Soldier must have served, or had the opportunity to serve, the time specified in the criterion. For example, to be included in the 6-month attrition criterion, a Soldier must have completed 6 months of service, or, if separated prior to 6 months, would have had to be eligible to serve at least 6 months had s/he stayed in the Army. As of the 31 December 2000 update, nearly all of the Regular Army Soldiers in the database had served long enough to establish a value for 18-month attrition.

Analysis Datasets

The AIM Grand Research Database includes data collected under several administration conditions. It also includes responses from members of the various Army components (i.e., Regular, Reserve, and Guard). Because of the breadth and variability of the database, investigators must use great care in selecting portions of the database for analyses. To minimize this problem and decrease the size of the files, we created three datasets with specific analysis objectives. These are described in Table 2.2.

In all the AIM datasets released for analysis, individual identification information has been removed (SSN and name). For a detailed description of the variables, see the *AIM Grand Research Database Codebook* (Barnes, 2001b).

Table 2.2. AIM Grand Research Database Analysis Dataset Descriptions

Database Title	Objective	Selection Criteria	No. Records
Attrition Research Dataset AIM_RSCH.SAV	Matching attrition data; retain only AIM data from original form; no Faking Study participants	Matched to EMF Not Faking Study	22,666
Faking Study Dataset AIM_FAKE.SAV	Faking Study – All data fields	Faking Study only	2,768
Alternate Forms Dataset AIM_ALTF.SAV	Original AIM and Alternate Form – all data fields	Alternate Form Data Present	13,430

Air Force AIM Research Database

Intrigued by preliminary AIM findings provided by ARI (White & Young, 1997), the Air Force became interested in the AIM as an attrition screen. During Fiscal Year (FY) 1998, the Air Force began a large-scale data collection using the AIM; most airmen entering basic training during that year completed the AIM. In this data collection, the Air Force used the original AIM. The Air Force provided its AIM database, which included the raw AIM data, information extracted from the Air Force MIRS extract, and the Air Force's Basic Military training (BMT) databases, to ARI.

The Air Force AIM Research Database contains records from 19,372 airmen who took the AIM during their first week of basic training at Lackland Air Force Base, Texas between October 1997 and September 1998. No data were available for the months of April or August, and the numbers of airmen tested were lower in March and May than in the other months. As a result of problems with missing and erroneous SSNs, 2,818 records could not be matched with MIRS and BMT databases.

We added separation information to the Air Force AIM Research Database first using updated data provided in June 1999 and then data updated through December 2000. ARI provided a list of SSNs to Air Force representatives who matched them to official Air Force separation records. If a match was made, the Air Force provided the date of separation (month and year). No additional information about the separation (e.g., separation program designator) was provided. The length of service and attrition criteria were computed in the same manner as for the Army AIM research sample. However, since only month and year information was provided, the calculations were performed using "months" as the basic unit of time, as opposed to "days," as used for the Army sample.

The Air Force research data provided the most mature attrition criteria. All members had served, or had the opportunity to serve, 24 months. Attrition at 36 months could be calculated for 87% of the sample.

As with the Army AIM databases, individual identification information was removed. For a detailed description of the variables, see the *AIM Air Force Research Database Codebook* (Barnes, 2001a).

Army AIM Operational Database

Use of the AIM as an adaptability-screening tool began in February 2000 with the implementation of the Army GED Plus program. The goal of the program is to recruit high aptitude youth who do not have a traditional high school diploma but are more likely to complete their enlistment obligation. The criteria to qualify for enlistment under GED Plus are stringent – Armed Forces Qualification Test (AFQT) Category I-IIIA, no moral or drug/alcohol enlistment waivers, individual left high school of own choice and is ineligible to return because of age, and a score in the top 75% on the AIM (based on norms obtained from the Army research sample). The program was implemented in accordance with an experimental design created by the RAND Corporation. During the RAND experiment, those who access under GED Plus are not counted against the Department of Defense's 10% cap on non-high school graduate accessions. In addition to GED Plus participants, many Army applicants whose educational credential placed them in Tier 2 or Tier 3 took the AIM. However, the AIM is not being used to determine enlistment eligibility outside of the GED Plus program.

As part of their GED Plus evaluation efforts, RAND (with assistance from ARI) built and maintains a database of Army GED Plus and other recruits who have entered the Army without a high school diploma. The database includes information that allows analysts to track the status of a recruit from the moment he or she entered the Delayed Entry Program (DEP). As of 31 March 2001, the database included records for 20,312 recruits. Of these, 6,699 were participants in the GED Plus Program, and all but two had AIM scores. An additional 3,965 recruits had AIM scores on record, although not as enlistees in the GED Plus Program.

HumRRO and ARI added attrition criteria and related variables from the MIRS extract and the EMF to create an operational database compatible with the Army AIM Research Database. The Army AIM Operational Database contains records for 5,832 Regular Army accessions into the GED Plus program. Although the first accessions entered service during February 2000, the attrition criteria were not mature for a large segment of the population in the database used for analyses documented in this report. For example, 32% of the population had not completed (or had the opportunity to complete) 6 months of service as of 30 June 2001. Nine-month attrition could not be calculated for half of the population.

CHAPTER 3. PRELIMINARY AIM VALIDATION BASED ON GED PLUS PROGRAM DATA

Dan J. Putka and Rodney A. McCloy
Human Resources Research Organization

The preliminary operational validation of AIM reported in this chapter provided the first indication that AIM was performing much differently in an operational setting (under GED Plus) than what had previously been observed under research conditions. In the GED Plus program, the mean AIM Adaptability score was approximately 1 standard deviation higher than that of Tier 2 recruits in our research sample. This level of score elevation was much higher than we had seen in the faking experiments conducted under the AIM Pre-Implementation Research Program (e.g., the effect sizes of .1 for the "Fake Operational" and .4 for the "Fake Maximal" conditions). This, and other more recent findings, have led to the conclusion that our simulated applicant faking experiments do not mimic the response set of highly motivated applicants in an operational setting. Moreover, we have learned that the results from our past faking experiments simply do not generalize to the operational environment.

Since this chapter was written, we have continued updating the Operational AIM Database for GED Plus applicants and reexamined the relationship between AIM Adaptability scores and attrition. These more recent findings show that AIM is a valid predictor of initial entry training (e.g., 6-month – 12-month) attrition among those tested under GED Plus. However, the magnitude of AIM's validity against attrition is only about one-third of the magnitude that was observed among those in our research sample. These disappointing findings increased our interest in exploring new ways of scoring AIM in an effort to preserve its validity under operational conditions. The ongoing work in this area, which includes the applications of highly complex Item Response Theory (IRT) models for scoring AIM, is being performed by our research team at the University of Illinois (Dr. Michael Levine, Dr. Bruce Williams, and Dr. Fritz Drasgow). Their initial findings, based on the Army's AIM Research Database are reported in Chapters 7 and 8. The more recent and highly promising work of these researchers (outside the timeframe of this report) has utilized data from the updated operational database.

Past evaluations of the AIM's validity for predicting attrition have been conducted exclusively in a research context. The primary purpose of the present investigation was to examine the validity of the AIM for predicting Soldiers' 3-, 6-, and 9-month attrition status using operational AIM data gathered as part of the GED Plus program.

As part of the current investigation, we examined the predictive validity of both the Adaptability Composite and its individual component scales. By comparing models of attrition based on the current Adaptability Composite with models based on its component scales, one can assess whether giving the most predictive components more weight when calculating an overall Adaptability score enhances validity.

In addition to examining the validity of the operational AIM, a secondary purpose of the present investigation was to examine the potential utility (e.g., hit rates, false acceptance rates) of implementing different cut scores in terms of decision quality. We contrasted the effects on the utility of the AIM of implementing cut scores at the 10th, 15th, and 25th percentiles with those achieved by the current cut score, which effectively screened out approximately 3.5% of the recruits who applied to the GED Plus program.[1]

Method

Participants

Data from 6,610 Soldiers who participated in the GED Plus program and who had Adaptability Composite scores of 46 or higher were drawn from the AIM Operational Database (see Chapter 2).[2] This sample of Soldiers primarily comprised Educational Tier 2 Soldiers (i.e., Soldiers holding an educational credential other than a high school diploma, such as a GED [a certificate of General Education Development]) (n = 5,897), with a smaller number of Educational Tier 3 Soldiers (i.e., Soldiers with no education credential) (n = 813). The database contained no Educational Tier 1 Soldiers (i.e., high school diploma graduates). The sample was primarily male (n = 4,936) and contained a number of Soldiers for whom gender information was missing (n = 931).

Some Soldiers in this sample were excluded from subsequent analyses because of missing or invalid attrition data. For example, depending on the attrition criterion considered, some Soldiers had not been with the Army long enough to have valid attrition data (e.g., a Soldier would have had to be with the Army at least 3 months to have valid 3-month attrition data). As a result of missing and invalid data, the effective sample sizes for the validation and utility analyses conducted in this investigation were reduced. Table 3.1 presents the sample sizes available for each set of analyses, as well as attrition rates for these samples at 3, 6, and 9 months. For purposes of contrast, Table 3.1 also includes the attrition rates of Tier 2 Soldiers from the research sample examined in Chapter 5 of this report. Attrition data for Soldiers in this Tier 2 research sample were drawn from the AIM Attrition Research Dataset.[3]

Not all Soldiers in the AIM Operational Database had a full set of Adaptability scores (composite and scales). Differences in sample sizes (n with the Adaptability Composite, n with all Adaptability scales) arose as a result of drawing composite and scale scores from two separate data sources. Despite the sample size differences, the attrition rates within these two sets of Soldiers were very similar. Because the AIM was not being used operationally in the Tier 2 research sample,

---

[1] The current AIM cut score was based on research data and was set at about the 25th percentile. Also, note that the screen out rate of 3.5% reported here is an approximation. Any Tier 3 applicants who failed the AIM under GED Plus would not be in the operational AIM database. Thus, we cannot know with certainty exactly what percentage of Soldiers is being screened out by the AIM. ARI expects that the actual screen out rate is not likely to be much different from what is reported here (Mark Young, personal communication, September 28, 2001).

[2] Eighty-seven Soldiers from the AIM Operational Database who were identified as having participated in the GED Plus program had Adaptability scores below the operational cut. These solders were excluded from all analyses. It is unclear why Soldiers who scored below the cut score accessed into the Army. One reason could be an issue of re-testing: Perhaps these low scores represent Soldiers' initial attempts at taking the AIM, and their later passing score was not recorded.

[3] The AIM Attrition Research Dataset is a subset of the AIM Grand Research Database and contains only Regular Army Soldiers who completed the AIM under non-faking conditions and for whom attrition data were available (see Chapter 2).

attrition rates for this group should be higher than those for the operational sample if the AIM is successfully screening out Soldiers who are more likely to attrit. Comparing the attrition rate in the Tier 2 research sample with the rate from the operational sample reveals little to no difference for the 6-month and 9-month samples. In the 3-month sample, however, the attrition rate is 9.7% lower in the operational sample compared to the research sample (1.5 percentage point difference). Note that the 95% confidence interval surrounding the 3-month attrition rate in the research sample is 14.1 to 16.7; thus, the observed difference between the operational and research sample is significant. These findings suggest that the current implementation of the Adaptability Composite reduces the 3-month attrition rate but has little effect on 6- or 9-month attrition.

Table 3.1. Sample Sizes and Attrition Rates

Sample / Predictors	3-Month Sample		6-Month Sample		9-Month Sample	
	n	Attrition Rate	*n*	Attrition Rate	*n*	Attrition Rate
Operational Sample						
Adaptability Composite	4,879	13.9	3,461	22.3	1,501	24.8
Adaptability Scales	4,129	14.1	2,976	22.4	1,271	24.1
Tier 2 Research Sample	3,181	15.4	3,180	22.3	3,180	25.0

Measures

As mentioned in the introduction, the only AIM variables examined in this investigation were the Adaptability Composite, its component scales, and a validity scale. The criteria examined in this investigation were Soldiers' 3-, 6-, and 9-month attrition status.

Analyses

Correlations

Several steps were taken to evaluate the validity of the AIM against the attrition criteria. First, we computed zero-order point-biserial correlations between each AIM variable and Soldiers' attrition status at 3, 6, and 9 months in the operational sample. Given that Soldiers in the present investigation were selected using the AIM, corrections for direct range restriction on the predictor were made to the raw coefficients (Guion, 1998) using variances from the research sample of AIM scores drawn from Tier 2 Soldiers in the AIM Attrition Research Dataset. In these and all subsequent analyses, Soldiers' attrition status was coded as 1 (attrit) or 0 (nonattrit).

Logistic Regressions

In addition to examining point-biserial correlations, we used logistic regression analyses to evaluate the efficacy of the Adaptability Composite as a predictor of attrition relative to using its component scales as predictors (i.e., entering each scale into the logistic model separately). All AIM variables were standardized across the operational sample prior to estimating the logistic regression models. Standardizing the AIM variables facilitated the interpretation of their corresponding conditional odds ratios from the logistic regression analysis. The conditional odds ratio for a predictor is formed by raising the mathematical quantity e to the power of the standardized partial beta weight of that predictor. In the present analyses, conditional odds ratios reflect the change in odds of a Soldier attriting, given an increase of one unit on the predictor variable of interest while holding all other predictors constant. Thus, a conditional odds ratio of 1.1 for a predictor would indicate that for every 1 unit increase on that predictor, a Soldier is 1.1 times as likely to attrit (relative to a Soldier at the next lowest unit on that predictor), holding all other predictors constant.

Several indexes were used to assess the validity of each logistic regression model examined. First, we generated a point-biserial correlation between the predicted probability of a Soldier's attrition and the Soldier's actual attrition status. We also calculated Cohen's d (effect size) index.[4] Cohen's d reflects the standardized mean difference between attritees and nonattritees in terms of their predicted probabilities of attriting as specified by a given model. We also calculated Negelkerke's R^2 for each model, which estimates the proportion of variance in Soldiers' likelihood of attrition that can be accounted for by predictors in the model of interest (Negelkerke, 1991).[5]

Several indexes from Signal Detection Theory (SDT) were also employed to evaluate the quality of each model. First, ROC (Receiver Operating Characteristic) curves were generated. In the context of examining attrition, ROC curves display the tradeoff between the hit rate (i.e., the proportion of attriting Soldiers who tested positive for attrition) and the false positive rate (i.e., the proportion of nonattriting Soldiers who tested positive for attrition) for each possible cut score for a given "test for attrition."[6] Each point on a ROC curve corresponds to a specific cut score on the diagnostic test (here the test for attrition). Each cut score has an associated hit rate (a y-coordinate) and false positive rate (an x-coordinate) (see Figure 3.1 for a sample ROC curve). Following the trace of a ROC curve from left to right reveals the tradeoff that *increasing the cut score* on a given test for attrition would have on these two proportions.

One particularly useful piece of information resulting from the generation of ROC curves is the area under the ROC curve (*AUC*), which here provides an indicator of the accuracy with which a given model predicts a Soldier's attrition status (Hanley & McNeil, 1982). Specifically, the *AUC* index for a given model reflects the expected proportion of times that an attritee would score higher than a nonattritee on the given attrition composite if a pair of Soldiers (one attritee and one nonattritee) was repeatedly selected at random from each group. For example, an *AUC*

[4] Although the point-biserial correlation and Cohen's d provide the same type of information regarding the predictive efficacy of a model, both statistics are presented for ease of comparison to past and future work.
[5] Negelkerke's R^2 was used as opposed to a more traditional Cox-Snell R^2 because the former has the desirable property of ranging from 0 to 1. This makes its interpretation more similar to the R^2 reported in traditional multiple regression.
[6] In the context of this investigation, coding attrition status as 1 (attrit) and 0 (nonattrit) resulted in predicted probabilities of attrition that correlated negatively with the AIM. Because many models consisted of multiple AIM scales, a Soldier's predicted probability of attriting served as the Soldier's score on the "test for attrition." This specification facilitates the use of the traditional medical SDT model, where attrition can be viewed as the "disease" and a reverse-scored AIM Composite (i.e., the test for attrition) as the test to detect the disease. One therefore can discuss cut scores on the "test for attrition" variable (i.e., the predicted probabilities resulting from a particular logistic model) or cut scores on the raw AIM variables. To clarify, higher cut scores on the test of attrition correspond to lower cut scores on the AIM, effectively screening out *fewer* Soldiers. Thus, screening out the top 10% of scorers on a test for attrition is the same as screening out the bottom 10% of scorers on the AIM.

value of .70 means that attritees would be expected to score higher than nonattritees on the attrition composite 70% of the time. *AUC* values range between .50 (indicating equal probability of attritees and nonattritees scoring higher) and 1.0 (indicating that attritees would always score higher than nonattritees). An additional benefit of *AUC* values is that they have a theoretical sampling distribution, which allows confidence intervals to be calculated, thus providing an inferential means to compare the accuracy of two or more models.[7]

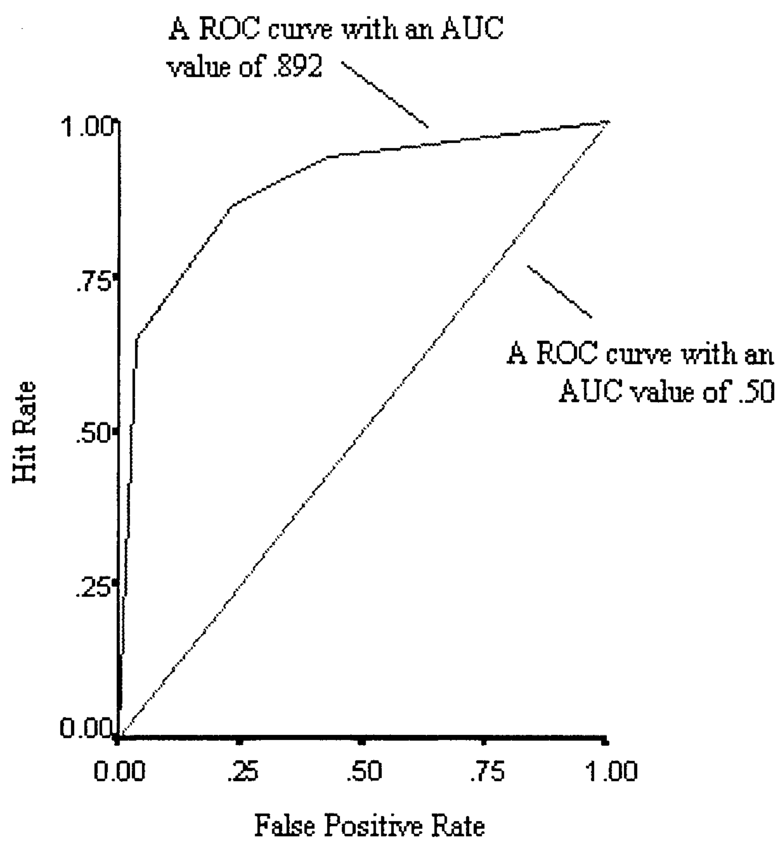


Figure 3.1. Sample ROC curve.

Evaluating the Utility of Different Models: Cut Score Analyses

Although examining the utility of the operational AIM can be achieved by comparing attrition rates from the operational samples to those from the research sample (see Table 3.1), such a comparison does not reveal the potential utility of the AIM using different cut scores.[8]

[7] Inferential comparisons between *AUC* values of different models in this investigation are conservative (i.e., estimated confidence intervals are wider then what they would likely be in reality) because the method used to generate standard errors assumes that the models being compared were estimated on independent samples of data (Hanley & McNeil, 1982). In this investigation, we compared *AUC* values for alternative AIM models using a single sample, thus violating the independence assumption.

[8] In the present investigation, "utility" refers to the expected quality of the decisions made for a given cut score on the predictor composites examined by each model, rather than monetary savings estimated to result from the use of the AIM as a screening tool.

The predictive value of a positive test (i.e., PVPT—the proportion of Soldiers who attrited, given that they tested positive for attrition based on a particular cut score) and false acceptance rate (i.e., the proportion of Soldiers who attrited, given they tested negative for attrition based on a particular cut score) were calculated for each AIM model under a variety of potential cut scores to examine the utility of the models for identifying Soldiers who were most likely to attrit at 3, 6, and 9 months. In this investigation, one can also interpret the false acceptance rate as the expected attrition rate should the cut score examined be implemented. As part of the present investigation, three cut scores were examined. Given the operational use of the AIM resulted in an approximate screen-out rate of only 3.5%, these cut scores were targeted to evaluate the utility of the AIM if approximately 10%, 15%, or 25% (goal of original cut) of applicants were screened out on the basis of their AIM scores.

Results

Descriptives

Means and standard deviations of each AIM variable for the operational sample are presented in Table 3.2. In addition to this information, means and standard deviations of AIM variables from the sample of Tier 2 Soldiers in the AIM Attrition Research Dataset are provided for comparison. The elevation of means in the operational sample compared to the unrestricted Tier 2 research sample is likely caused by both response distortion on the part of applicants in the operational sample (i.e., faking), and the lack of direct range restriction in the unrestricted research sample (i.e., the lower bound on the Adaptability scores was less than 46). To attempt to tease apart the proportion of these differences due to response distortion, a restricted sample containing only Tier 2 Soldiers who scored 46 or above was drawn from the AIM Attrition Research Dataset (this reduced the research sample by approximately 25%). Limiting the restricted Tier 2 research sample to only Soldiers exceeding the current operational cut score made direct comparison of the means and standard deviations with the operational sample cleaner by eliminating the direct range restriction explanation for the differences between means. As Table 3.2 reveals, the mean Adaptability score in the operational sample was still substantially larger than the mean Adaptability score in the restricted research sample (about 0.85 standard deviations). This finding, coupled with the small change in mean validity scale (Scale G) scores in these two samples (about 0.15 standard deviation), suggests that a great degree of response distortion was occurring in the operational sample and the validity scale was not particularly effective at detecting it.[9]

[9] ARI suggested an alternative reason for the apparent elevation in applicant sample AIM scores relative to the research sample AIM scores. Specifically, they hypothesized that Soldiers in the research sample experienced "the stress of civilians entering the Army and being emotionally overwhelmed," thus leading to depression of their AIM scores (Mark Young, personal communication, September 28, 2001). Another potential reason for depression of scores in the research sample is that Soldiers' AIM scores were not tied to tangible personal outcomes (i.e., their acceptance into the Army); thus, Soldiers in the research sample may have lacked the incentive to take the AIM seriously. Such a lack of incentive may have resulted in higher proportions of random responding which could in turn lead to lower scores in the research sample relative to the operational sample.

Table 3.2. Means and Standard Deviations of AIM Variables

Predictor	Operational Sample			Tier 2 Research Sample			
				Unrestricted[a]		Restricted[b]	
	n	*M*	*SD*	*M*	*SD*	*M*	*SD*
Adaptability Composite	6,610	64.64	7.11	53.89	11.08	58.41	7.58
Scale A	5,707	26.36	3.53	21.71	5.38	23.64	4.06
Scale B	5,707	26.98	3.73	22.76	5.01	24.53	3.86
Scale C	5,707	11.39	2.43	9.42	2.96	10.24	2.54
Scale G	5,706	1.59	1.84	1.22	1.35	1.36	1.40

Note. [a]The unrestricted Tier 2 research sample consists of all Tier 2 Soldiers from the AIM Attrition Research Dataset who had Adaptability scores. [b]The restricted Tier 2 research sample consists of only those Tier 2 Soldiers from the AIM Attrition Research Dataset who had Adaptability scores of 46 or greater.

Validity of the AIM

Table 3.3 presents (a) point-biserial correlations between the AIM variables and attrition status for the operational GED Plus sample, and (b) Cohen's *d* values, which reflect the standardized mean difference between attritees' and nonattritees' scores on each AIM variable. Positive *d* values indicate that nonattritees scored higher than attritees on the AIM variable of interest.

Based on Table 3.3, AIM Scale C was the only component of the Adaptability Composite that was significantly negatively related to Soldiers' attrition status at all three times examined (i.e., 3, 6, and 9 months). AIM Scale B exhibited positive correlations (statistically nonsignificant) with both 3- and 6- month attrition status, and zero correlation with 9-month attrition status. Although both the Adaptability Composite and AIM Scale A were negatively correlated with attrition status for all months examined, these correlations failed to reach statistical significance.

The present findings suggest that forming an AIM composite that is weighted more heavily by Scale C (under the current scoring it receives the least weight) may help to improve its validity for predicting early career attrition. To evaluate the efficacy of such a strategy, Table 3.4 presents the results of logistic regression analyses that fit attrition models based on the Adaptability Composite as a predictor, and models that fit the Adaptability scales as separate predictors.

Results of fitting the logistic regression models to the data revealed that the Adaptability composite failed to provide a significantly better fit to the attrition data than the null model (i.e., just the intercept, no predictors at all) for all attrition criteria examined. However, when the individual components of the Adaptability Composite were entered separately, the resulting models provided significantly improved fit over and above the null model for all attrition criteria with the exception of 9-month attrition status. In addition to significant model fit, the Adaptability-component models resulted in significantly greater point-biserial correlations and *d* values compared to the Adaptability Composite models. With regard to the individual components of the Adaptability Composite, Scale C was the strongest predictor of early career attrition status, holding constant the level of both Scale A and Scale B.

Table 3.3. Point-Biserial Correlations between AIM Variables and Attrition Status for the Operational Sample

Predictor	*n*	*r*	$M_{Attritees}$	$M_{Nonattritees}$	*d*
		3-Month Sample			
Adaptability Composite	4,879	-.020 (-.013)	64.49	64.75	0.037
Scale A	4,129	-.014 (-.009)	26.25	26.34	0.026
Scale B	4,129	.024 (.018)	27.17	26.98	-0.050
Scale C	4,129	-.079 (-.065***)	11.07	11.52	0.187
Scale G	4,128	.020	1.70	1.59	-0.057
		6-Month Sample			
Adaptability Composite	3,461	-.025 (-.016)	64.58	64.85	0.038
Scale A	2,976	-.011 (-.007)	26.32	26.37	0.016
Scale B	2,976	.020 (.015)	27.07	26.94	-0.035
Scale C	2,976	-.082 (-.067***)	11.18	11.56	0.160
Scale G	2,975	.010	1.63	1.58	-0.025
		9-Month Sample			
Adaptability Composite	1,501	-.061 (-.039)	64.74	65.36	0.090
Scale A	1,271	-.040 (-.026)	26.42	26.63	0.060
Scale B	1,271	.000	27.00	27.01	0.001
Scale C	1,271	-.083 (-.068**)	11.27	11.65	0.159
Scale D	1,271	.011	1.56	1.52	-0.025

Note. Attrition status was coded as 0 (nonattrit) or 1 (attrit). Correlations outside the parentheses have been corrected for range restriction on the predictor. Raw correlations appear inside the parentheses. ** $p < .01$, *** $p < .001$.

Although improvement was apparent with the Adaptability-component models, neither the component models nor the Adaptability Composite models were particularly effective at predicting variation in Soldiers' attrition status. For example, although the Adaptability-component models provided significantly improved fit relative to the null model, the magnitude of the relationship with attrition was minimal, as no point-biserial correlation exceeded .08 (note that no correlation for the Adaptability Composite exceeded .04). These values reflect that, at best, the Adaptability models accounted for less than 1% of the variance in Soldiers' likelihood of attrition.

Another way to interpret the validity of the AIM models for differentiating between attritees and nonattritees is to consider the *AUC* values associated with each model (Table 3.4). Based on the *AUC* values for the Adaptability Composite models, attritees would be expected to receive lower Adaptability scores than nonattritees only 51.3% to 53.3% of the time (reverse scoring the model scale for ease of interpretation). Such values indicate that the Adaptability Composite alone is not distinguishing particularly well between attritees and nonattritees, as an *AUC* value of .50 indicates pure chance. With regard to the *AUC* values for the Adaptability-component models, in all cases (with the exception of the 9-month attrition) these values were significantly greater than the *AUC* value from their corresponding Adaptability Composite model. Although these differences were statistically significant, the *AUC* values for the Adaptability-component models suggest that attritees would only be expected to receive higher Composite scores (higher in this case means more likely to attrit) than nonattritees between 53.7% and 55.0% of the time. Given that constructing a model based on the flip of a balanced coin would result in an *AUC* estimate of .50, the Adaptability-component models do not appear to provide much improvement over the Adaptability Composite models.

Table 3.4. Comparison of the Validity of Different AIM Combinations for Predicting Attrition Status

Model Statistics	3-Month Sample		6-Month Sample		9-Month Sample	
	Composite	Scales	Composite	Scales	Composite	Scales
			Conditional Odds Ratios			
Adaptability Composite	0.962	.	0.961	.	0.910	.
Scale A	.	0.996	.	1.008	.	0.964
Scale B	.	1.101*	.	1.073	.	1.045
Scale C	.	0.812***	.	0.835***	.	0.851*
			Model-Level Validity			
Model χ^2	0.81	21.39***	0.86	15.71**	2.24	6.27
R	.013	.076	.016	.074	.038	.073
d	-0.038	-0.218	-0.038	-0.178	-0.088	-0.172
Nagelkerke R^2	.000	.009	.000	.008	.002	.007
AUC	.515	.550	.513	.545	.533	.537
AUC 95% C.I.	(.490,.540)	(.525,.576)	(.489,.538)	(.520,.569)	(.497,.569)	(.500,.575)

Note. Attrition status was coded as 1 (attrit) or 0 (nonattrit). Standardized partial beta weights for each predictor can be obtained by taking the natural log of a predictor's conditional odds ratio (i.e., β = ln(conditional odds ratio). The model χ^2 tested the given model against a null model with no predictors. The χ^2 test was based on 1 degree of freedom for the Composite models and 3 degrees of freedom for the scale models. ** $p < .01$, *** $p < .001$.

Utility of Various AIM Cut Scores

Although the AIM appears to be accounting for very little variation in the attrition criteria, this does not necessarily imply that the variance it accounts for is not meaningful. Indeed, even small amounts of validity can translate into substantial gains in terms of utility (Hunter & Schmidt, 1990). For example, recall the results presented in Table 3.1. Using the Adaptability Composite to screen out recruits in the operational sample reduced 3-month attrition by approximately 9.7% relative to the Tier 2 attrition base rate. This finding emerged despite the fact that the point-biserial correlation between Soldiers' 3-month attrition status and their Adaptability Composite score was only -.013. Thus, low validity does not necessarily equate to low utility. To further examine the potential utility of the AIM for screening out prospective recruits, three other potential cut scores were considered. These revised cuts correspond to making 10%, 15%, and 25% operational cuts, as opposed to the 3.5% cut realized in the current operational sample. Table 3.5 presents the results of these cut score analyses.

Based on the results presented in Table 3.5, there appears little evidence to suggest that raising the cut score on the AIM Composite would result in a greater reduction of attrition than is occurring with the current cut score (which in the operational sample equates to about a 3.5% cut). Specifically, the maximum percentage point decrease in attrition rate under different cut scores relative to the current cut was 0.2 for 3-month attrition, 0.1 for 6-month attrition, and 0.5 for 9-month attrition. Comparing the PVPT for the current cut to those of the other cut scores is not possible because no one who was predicted to attrit based on the current cut was screened in to the Army. Nevertheless, the PVPT values for the other cuts were low. With regard to the utility of Composites based on the Adaptability-component models, a similar pattern emerged. The maximum percentage point decrease in attrition rate under different cuts relative to the current cut was slightly higher, however, namely 0.9 for 3-month attrition, 0.8 for 6-month attrition, and 1.2 for 9-month attrition. Another (perhaps more salient) difference between the Adaptability Composite and Adaptability-component models regarded the PVPT values. Specifically, PVPT values tended to be higher for the Adaptability-component models than for the Adaptability Composite models, and these differences increased as more mature attrition samples were considered.

Table 3.5. Comparison of Different AIM Combinations for Predicting Tier 2 Soldiers' 18-Month Attrition Status

Model Statistics	3-Month Sample		6-Month Sample		9-Month Sample	
	Composite	Scales	Composite	Scales	Composite	Scales
Estimated 10% Cut						
PVPT	19.2	19.8	24.2	30.7	24.7	38.0
False Acceptance Rate	13.8	13.7	22.2	21.9	24.8	23.3
Estimated 15% Cut						
PVPT	18.8	19.2	23.4	29.5	26.8	33.8
False Acceptance Rate	13.7	13.5	22.2	21.6	24.6	22.9
Estimated 25% Cut						
PVPT	14.3	17.5	22.3	25.6	26.6	27.7
False Acceptance Rate	13.8	13.2	22.3	21.6	24.3	23.2
Current Cut Attrition Rate	13.9	14.1	22.3	22.4	24.8	24.1
Tier 2 Research Sample Attrition Rate	15.4		22.3		25.0	

Note. PVPT refers to the predictive value of a positive test (i.e., the proportion of Soldiers who attrited, given that they tested positive for attrition based on the given cut). Current cut attrition rates are based on the estimated screen-out rate of approximately 3.5%. Research sample attrition rates are based on attrition rates observed among Tier 2 recruits from the AIM Attrition Research Dataset.

Discussion

Although the current investigation found low levels of validity with regard to the ability of the AIM Adaptability Composite to predict 3-, 6-, and 9-month attrition, these findings need qualification. First, although little validity evidence emerged for the Adaptability Composite, AIM Scale C (one of three components of the Adaptability Composite) consistently showed significant relationships to Soldiers' attrition status at 3, 6, and 9 months. These findings, as well as the relative lack of validity found for the other components of the Adaptability Composite, suggest that the validity of the Composite may be improved if more weight (at least equal weight) were given to Scale C. Currently, AIM Scales A and B can contribute up to 40 and 38 points, respectively to the Adaptability Composite, while Scale C can contribute only up to 16 points. The results of subsequent logistic regression analyses also were consistent with the suggestion that Scale C should be given more weight, at least if the Army's concern is creating a more valid predictor of 3-, 6-, and 9-month attrition status among Tier 2 and Tier 3 Soldiers.

A second qualifying statement regarding the lack of validity found for the AIM Adaptability Composite regards its utility for reducing 3-, 6-, and 9-month attrition. Although low levels of validity were found, based on the current cut score, the 3-month attrition rate was reduced by 1.5 percentage points (See Table 3.5; 15.4 – 13.9 = 1.5) through implementing the AIM Adaptability Composite operationally. Implementing the AIM with the current cut score however, or any of the alternative cut scores examined, did not appear to substantially decrease the expected rate of attrition at 6 months or 9 months relative to their respective base rates from the Tier 2 Soldiers in the AIM Attrition Research Dataset.

Despite the above qualifying statements, one may still question the ability of the operational AIM to predict attrition, particularly among Tier 2 and Tier 3 recruits. Nevertheless, caution should be taken before dismissing the AIM in general as an ineffective predictor of Tier 2 and Tier 3 attrition. This investigation examined the Adaptability Composite as it is currently being scored, and allowed each component of the composite to be weighted by its partial beta weight. The results suggest that some validity and utility may be gained by reconsidering how the components of the composite are combined to yield the Adaptability score. In addition, alternative scoring routines for the components are being examined, and the results of those analyses are presented in Chapter 7 of this report. Between reweighting the components of the Adaptability Composite, and finding optimal scoring routines for those components, a reassessment of the validity and utility of the AIM using such new scoring techniques should be conducted to see if substantial improvement results.

Lastly, although the AIM appears to lack validity for predicting 3-, 6-, and 9-month attrition (as its components are currently weighted and scored), this analysis cannot speak to the validity of the AIM for predicting more mature attrition criteria. For example, although the AIM may be relatively ineffective at predicting attrition early in Soldiers' tenure, it may be more predictive of attrition later on in their development (e.g., 18–36 months).

CHAPTER 4. AIM ALTERNATE FORM SELECTION AND SCALING

Rodney A. McCloy and Carol E. George
Human Resources Research Organization

Charlie L. Reeve
Purdue University

When our Post-Implementation Program began in September 2000, the original AIM form was being used to screen applicants for the GED Plus Program, but no alternate forms were ready for implementation. Although alternate forms had previously been created under the AIM Pre-Implementation Research Program (1998–1999), preliminary analyses indicated that these forms were very similar to, but not psychometrically "equivalent" to the original form. The effort reported in this chapter was conducted to move the existing alternate forms closer to implementation. This work relied solely on the AIM research database because the alternate forms had never been administered in an operational setting.

The effort described in this chapter was carried out under two assumptions: (a) That the AIM Adaptability Composite would be viable as an operational attrition screen for the GED Plus Program, and (b) that findings from the research database would generalize to the operational setting. It turned out that neither of these assumptions was tenable. We eventually learned (from Chapter 3) that the AIM Adaptability Composite as originally scored performed poorly in the operational setting and that its performance in a research setting did not generalize well to its performance in an operational context.

The work reported here serves to highlight the difficulties of developing and evaluating alternate forms for motivational measures (like AIM) which use a complex forced-choice format. Unlike cognitive measures such as the Armed Services Vocational Aptitude Battery (ASVAB), there are no well-established procedures for developing equivalent forms for AIM-like instruments, and it is uncertain as to whether true equivalence can be achieved with such measures. Similarly, there is no universal agreement on the appropriate procedures for evaluating the correspondence between alternate forms on these types of instruments.

Background

Recruits who do not possess a high school diploma are less likely to complete their first term of service than are high school diploma graduates (Trent & Laurence, 1993). Previous research on the AIM (Heggestad, Young, Strickland, & Rumsey, 1999; Young, Heggestad, Rumsey, & White, 2000) demonstrated that higher scores on the AIM Adaptability Composite were associated with lower rates of first-term attrition. Based in part on these results, the Army began administering the AIM operationally to nongraduates in February 2000 as part of the GED Plus program. The Army denies enlistment to some GED Plus applicants scoring below the

Adaptability Composite cut score. However, in many cases a Tier 2 (but not a Tier 3) applicant may be able to enter the Army even if he/she was not accepted under GED Plus.

Alternate Forms

The operational administration of AIM raises the issue of test compromise. To minimize the potential for compromise, ARI contracted with HumRRO to construct alternate forms of AIM (Heggestad, Young, et al., 1999). As discussed in Chapter 2, HumRRO developed two sets of two alternate forms: the *initial* alternate forms (identified as Forms A and B) and the *revised* alternate forms (identified as Forms A* and B*).

The initial alternate forms were administered to 2,709 Soldiers (approximately 33% of whom were not Regular Army) between 31 October 1998 and 17 January 1999. The item structure of 3 of 26 items on Forms A and B differed slightly from the original form. All three items involved a stem from Scale G. The means for all scales except Scale G were higher for Forms A and B than for the original form.

The revised alternate forms were administered to 10,576 Soldiers (approximately 35% of whom were not Regular Army) between 23 January 1999 and 18 April 1999. The revised forms were developed to improve various psychometric characteristics of the original alternate forms.

Analysis Goals

Given two sets of alternate forms and the need to have more than a single operational test form, this task addressed two primary goals: (a) to select a set of alternate forms to be scaled to the original form, and (b) to effect the scaling. The analyses performed to meet these goals and the results thereof are discussed in subsequent sections.

Selection of AIM Alternate Forms to Be Scaled to the Original Form

HumRRO conducted several comparative analyses to determine whether the initial or revised alternate forms would be better candidates for scaling. These included analyses that compared Forms A and B and Forms A* and B* on (a) the rank-order correlations between scale scores from each form and the original AIM, (b) the extent to which they yield decisions consistent with the original AIM, and (c) their ability to predict attrition. Although both sets of forms were ultimately scaled to the original AIM, the process of evaluating the two sets is still informative and will be described in some detail.

The decision consistency approach to evaluating the AIM forms merits a bit more explanation. Here, the task of selecting alternate forms focuses on identifying those forms with the greatest functional similarity to the original AIM: that is, the alternate forms that yielded examinee decisions (i.e., reject/accept) that were most consistent with those from the original AIM. Ideally, this approach would require us to determine equivalent scores across forms—a task that was to follow the identification of alternate forms but precede the analysis of decision consistency. Nevertheless, given that the AIM is used to make operational pass/fail decisions about individuals, decision consistency analyses were viewed as most relevant for the goal of identifying the appropriate alternate forms to scale to the original AIM. This approach focuses on the most crucial operational question: Which alternate forms result in the most similar reject/accept decisions *for individuals* in a given sample of examinees?

Method

Evaluation of the alternate forms of AIM involved three analyses. Two analyses concerned the consistency of the various forms. The first analysis focused on the consistency of the *rank-ordering* of respondents across forms. Rank-order correlations were calculated between the original AIM and the four alternate forms. Also, the mean change in rank-order across forms was calculated.

The second analysis addressed the consistency of *decisions* about respondents when using the various forms. We selected three cutoff scores (10^{th} percentile, 25^{th} percentile, 35^{th} percentile) on the Adaptability Composite distribution from the original form. We then identified corresponding Adaptability scores on the alternate forms. The corresponding scores for the alternate forms were those closest to the 10^{th}, 25^{th}, and 35^{th} percentiles of their respective distributions. After identifying the corresponding cutoff scores, we constructed frequency tables that classified individuals as scoring either below or at/above the cut score on each form. We converted the frequencies to proportions, and calculated statistics to address the decision consistency of the forms in question.

The third analysis moved from issues of consistency to predictive power. Here, the focus was the degree to which each form predicted first-term attrition during periods of 3, 6, 9, 12, and 18 months. Point-biserial correlations were calculated between (a) the Adaptability composite and its component scales and (b) a 1/0 dummy variable indexing attrition status (1 = attrition).

Sample

Data for these analyses were obtained from the Army AIM Grand Research Database (see Chapter 2 of this report). Samples sizes were as follows: 1,674 Soldiers had scores on the original form and Form A, 1,673 Soldiers had scores on the original form and Form B, and 7,160 Soldiers had scores on the original form and Forms A*/B*.

Results

Rank-Order Correlations

To examine consistency across the AIM forms, we calculated the rank-order correlation (i.e., Spearman's ρ) between (a) the ranks of Adaptability Composite and the ranks of each trait scale from the original form and (b) the ranks of the counterpart composite/scale from the alternate forms. Rank-order correlations were also calculated between the initial alternate forms and between the revised alternate forms. The results of this analysis appear in Table 4.1. The table indicates the following:

- Overall, the values of Spearman's ρ for the original form are a bit low, given that they may be interpreted as an estimate of alternate-forms reliability.

- The highest rank-order correlations were attained for the Adaptability Composite for all comparisons examined, ranging from .69 to .71.

- The lowest rank-order correlations were attained for Scale G for all comparisons examined, ranging from .33 to .37.

- For the Adaptability Composite and each scale (except Scale G), Forms A/B show equal or slightly higher rank-order correlations with most respective scales on the original form than do Forms A*/B*.

- For all scales but Scale G, the alternate forms yield higher rank-order correlations with each other than with the original form.

Table 4.1. Rank-Order Correlations (ρ) of Alternate Forms with the Original Form

	ρ Between Original Form and				ρ Between Alternate Forms	
Score	A	B	A*	B*	A/B	A*/B*
Adaptability	.71	.69	.69	.70	.80	.76
Scale A	.62	.63	.62	.64	.72	.69
Scale B	.62	.60	.58	.54	.67	.59
Scale C	.53	.55	.53	.53	.62	.57
Scale D	.47	.43	.49	.47	.60	.62
Scale E	.58	.59	.53	.55	.64	.61
Scale F	.60	.62	.61	.59	.68	.65
Scale G	.33	.33	.36	.37	.32	.37

Another analysis involving rank-orders across AIM forms involved calculating the mean difference between the examinees' percentile ranks across forms. The results appear in Table 4.2. The table indicates the following:

- Changes in percentile rank across forms are often considerable, with at least 50% of the examinees experiencing double-digit changes in percentile rank (the distributions of percentile ranks show evidence of moderate positive skew, although the means do not differ greatly from the medians).

- Mean percentile rank changes between the original form and the two sets of alternate forms are virtually identical.

- Forms A/B provide more consistent rank orderings of examinees (mean difference of 13.69) than do Forms A*/B* (mean difference of 15.32).

Table 4.2. Descriptive Statistics for Individual Percentile Rank Differences Between Forms

Form Pairs	*n*	Mean	Mdn	*SD*	Min	Max	Skew	Kurt
Original v. A	1,674	17.12	14	13.94	0	88	1.08	1.13
Original v. B	1,673	17.14	13	14.64	0	88	1.20	1.52
A v. B	1,675	13.69	11	12.14	0	85	1.49	3.13
Original v. A*	7,160	16.99	13	14.67	0	93	1.32	1.84
Original v. B*	7,160	17.15	14	14.47	0	91	1.23	1.55
A* v. B*	7,194	15.32	12	13.07	0	92	1.31	2.12

Note: Mean = Absolute percentile difference, which is calculated as $| \%ile_{form1} - \%ile_{form2} |$. Mdn=median; *SD*=standard deviation; Min=minimum; Max=maximum; Skew=skewness; Kurt=kurtosis.

Decision Consistency

The frequency distributions and associated proportions for the paired forms are given in Tables 4.3 through 4.8. Tables 4.3 through 4.5 present statistics from pairings of the original AIM with Forms A, B, A*, and B* with cut scores set at the 10th, 25th, and 35th percentiles, respectively.

Because the alternate forms will be considered on equal standing with the original form (following the scaling exercise), it is also important to examine the decision consistency across alternate forms. To examine this question, Tables 4.4 through 4.6 present statistics from pairings of the alternate forms within a set (i.e., A with B, A* with B*) with cut scores set at the 10th, 25th, and 35th percentiles, respectively.

Also provided in Tables 4.3 through 4.8 are the following statistics that describe decision consistency across the AIM forms:

- *Observed Consistency*—the proportion of consistent decisions across forms (i.e., reject/reject and accept/accept).

- *Chance Consistency*—the amount of consistency that would be expected by chance alone (i.e., that attained when the scores on the forms are statistically independent), which is defined as the sum of products of marginal proportions:

$$P_c = P_{1.}P_{.1} + P_{0.}P_{.0}$$

where P_c is chance consistency, $P_{1.}$ is the proportion of examinees exceeding the cut score (i.e., accepted) on the first form, $P_{.1}$ is the proportion of examinees exceeding the cut score on the second form, $P_{0.}$ is the proportion of examinees scoring below the cut score (i.e., rejected) on the first form, and $P_{.0}$ is the proportion of examinees scoring below the cut score on the second form.

- *Cohen's Kappa*—an index of decision consistency that "may be interpreted as the increase in decision consistency that the tests provide over chance expressed as a proportion of the maximum possible increase over chance consistency" (Crocker & Algina, 1986, p. 201). Kappa is calculated thus:

$$\kappa = \frac{P - P_c}{1 - P_c}$$

  where P is observed consistency and P_c is chance consistency.

- *Low/High 95% CI*—the lower and upper bounds of the 95% confidence interval around kappa. The interval is based upon the standard error of kappa—the square of which is listed in the tables as Var(k)—reported in Fleiss, Cohen, and Everitt (1969).

The consistency statistics presented in Tables 4.3 through 4.8 indicate the following:

- Observed consistency is substantial, ranging from 78% (original form with Form A*, cut score at 35th percentile [Table 4.5]) to 91% (Form A with the original form, cut score at 10th percentile [Table 4.3] and Form A with Form B, cut score at 10th percentile [Table 4.6]).

- Chance consistency varies substantially with placement of the cut score on the Adaptability Composite, averaging 83%, 64%, and 56% at the 10th, 25th, and 35th percentiles, respectively.

- Kappa increases with the cut score and is highest at all cut scores for Forms A and B (Tables 4.3 to 4.5).

- Forms A and B yield greater decision consistency with the original form than do Forms A* and B* (Tables 4.3 through 4.5), with comparisons of kappa at the two higher cut scores yielding statistically significant differences.

- Forms A and B yield similar degrees of decision consistency (with the original form) at the higher cut scores, with Form A yielding greater consistency at the 10th percentile.

- Forms A and B yield significantly greater decision consistency with one another than do Forms A* and B* (Tables 4.6 through 4.8); indeed, the best decision consistency would be obtained if the original form were dropped and Forms A and B were used in its place.

Table 4.3. Decision Consistency for Original Form with Alternate Forms: Cut Score = 10th Percentile

Decision Original Form	Decision Alternate Form Reject	Decision Alternate Form Accept	Proportions Reject	Proportions Accept	Observed Consistency	Chance Consistency	Kappa	Var (k)	Low 95% CI	High 95% CI
	Form A									
Reject	75	87	0.045	0.052	0.912	0.838	0.46	0.001424	0.38	0.53
Accept	60	1,452	0.036	0.867						
	Form B									
Reject	84	78	0.050	0.047	0.892	0.813	0.42	0.001259	0.35	0.49
Accept	103	1,408	0.062	0.842						
	Form A*									
Reject	284	364	0.040	0.051	0.903	0.839	0.40	0.000344	0.36	0.43
Accept	330	6,182	0.046	0.863						
	Form B*									
Reject	256	392	0.036	0.055	0.903	0.845	0.37	0.000357	0.33	0.41
Accept	304	6,208	0.042	0.867						

Table 4.4. Decision Consistency for Original Form with Alternate Forms: Cut Score = 25th Percentile

Decision Original Form	Decision Alternate Form Reject	Decision Alternate Form Accept	Proportions Reject	Proportions Accept	Observed Consistency	Chance Consistency	Kappa	Var (k)	Low 95% CI	High 95% CI
	Form A									
Reject	248	154	0.148	0.092	0.832	0.643	0.53	0.000606	0.48	0.58
Accept	128	1,144	0.076	0.683						
	Form B									
Reject	270	132	0.161	0.079	0.826	0.627	0.53	0.000574	0.49	0.58
Accept	159	1,112	0.095	0.665						
	Form A*									
Reject	949	714	0.133	0.100	0.807	0.647	0.45	0.000155	0.43	0.48
Accept	670	4,827	0.094	0.674						
	Form B*									
Reject	904	759	0.126	0.106	0.806	0.653	0.44	0.00016	0.42	0.46
Accept	633	4,864	0.088	0.679						

Table 4.5. Decision Consistency for Original Form with Alternate Forms: Cut Score = 35th Percentile

Decision Original Form	Decision Alternate Form Reject	Decision Alternate Form Accept	Proportions Reject	Proportions Accept	Observed Consistency	Chance Consistency	Kappa	Var (k)	Low 95% CI	High 95% CI
	Form A									
Reject	379	203	0.226	0.121	0.797	0.559	0.54	0.000476	0.50	0.58
Accept	136	956	0.081	0.571						
	Form B									
Reject	429	152	0.256	0.091	0.800	0.541	0.56	0.000444	0.52	0.61
Accept	182	910	0.109	0.544						
	Form A*									
Reject	1,563	735	0.218	0.103	0.776	0.557	0.49	0.000119	0.47	0.51
Accept	870	3,992	0.122	0.558						
	Form B*									
Reject	1,466	832	0.205	0.116	0.785	0.570	0.50	0.000122	0.48	0.52
Accept	710	4,152	0.099	0.580						

Table 4.6. Decision Consistency for Alternate AIM Forms: Cut Score = 10th Percentile

Decision A/A*	Decision B/B* Reject	Decision B/B* Accept	Proportions Reject	Proportions Accept	Observed Consistency	Chance Consistency	Kappa	Var (k)	Low 95% CI	High 95% CI
	A/B									
Reject	88	48	0.053	0.029	0.912	0.825	0.50	0.001270	0.43	0.57
Accept	99	1,440	0.059	0.860						
	A*/B*									
Reject	250	370	0.035	0.051	0.905	0.849	0.37	0.000363	0.33	0.41
Accept	315	6,259	0.044	0.870						

Table 4.7. Decision Consistency for Alternate AIM Forms: Cut Score = 25th Percentile

Decision A/A*	Decision B/B* Reject	Decision B/B* Accept	Proportions Reject	Proportions Accept	Observed Consistency	Chance Consistency	Kappa	Var (k)	Low 95% CI	High 95% CI
	A/B									
Reject	292	85	0.174	0.051	0.867	0.634	0.64	0.000489	0.59	0.68
Accept	137	1,161	0.082	0.693						
	A*/B*									
Reject	892	741	0.124	0.103	0.806	0.655	0.44	0.000161	0.41	0.46
Accept	657	4,904	0.091	0.682						

Table 4.8. Decision Consistency for Alternate AIM Forms: Cut Score = 35th Percentile

Decision A/A*	Decision B/B* Reject	Decision B/B* Accept	Proportions Reject	Proportions Accept	Observed Consistency	Chance Consistency	Kappa	Var (k)	Low 95% CI	High 95% CI
	A/B									
Reject	427	89	0.255	0.053	0.836	0.552	0.64	0.000394	0.60	0.67
Accept	185	974	0.110	0.581						
	A*/B*									
Reject	1,566	886	0.218	0.123	0.790	0.562	0.52	0.000115	0.50	0.54
Accept	626	4,116	0.087	0.572						

Predictive Validity

One other analysis that informed the recommendation of alternate AIM forms to scale to the original form entailed estimating the predictive validity of the alternate forms with attrition at various time points. Table 4.9 presents point-biserial correlations between (a) the Adaptability Composite and its three component scales (Scales A, B, and C) from the original form and from both sets of alternate forms, and (b) attrition at 3, 6, 9, 12, and 18 months. The correlations are based on listwise deletion of missing cases (data for the original form are based only on those Soldiers who also completed the alternate forms).

Table 4.9. Predictive Validity of the Adaptability Composite and Its Component Scales from the Original Form and the Initial and Revised Alternate Forms: Listwise Deletion

		Attrition (Months)				
Form	Score	3	6	9	12	18
Original[a]	Adaptability	-.14	-.17	-.14	-.14	-.14
	Scale A	-.13	-.16	-.13	-.13	-.12
	Scale B	-.09	-.10	-.09	-.08	-.10
	Scale C	-.12	-.15	-.13	-.13	-.13
Form A	Adaptability	-.15	-.18	-.18	-.18	-.17
	Scale A	-.13	-.16	-.15	-.15	-.14
	Scale B	-.11	-.13	-.13	-.12	-.14
	Scale C	-.13	-.17	-.16	-.16	-.16
Form B	Adaptability	-.13	-.15	-.14	-.14	-.15
	Scale A	-.14	-.15	-.14	-.14	-.14
	Scale B	-.08	-.09	-.08	-.08	-.08
	Scale C	-.10	-.12	-.12	-.12	-.14
Original[b]	Adaptability	-.12	-.13	-.13	-.13	-.13
	Scale A	-.12	-.12	-.12	-.12	-.11
	Scale B	-.06	-.06	-.07	-.07	-.07
	Scale C	-.11	-.13	-.14	-.13	-.13
Form A*	Adaptability	-.12	-.13	-.13	-.12	-.13
	Scale A	-.12	-.13	-.13	-.12	-.12
	Scale B	-.07	-.07	-.07	-.07	-.07
	Scale C	-.10	-.11	-.11	-.11	-.11
Form B*	Adaptability	-.11	-.12	-.12	-.11	-.11
	Scale A	-.11	-.12	-.12	-.11	-.11
	Scale B	-.05	-.06	-.06	-.06	-.06
	Scale C	-.09	-.10	-.11	-.10	-.10

Note: [a]n = 1,686; [b]n = 6,539.

Table 4.9 reveals that at most time points, Forms A/B are more predictive of attrition than are either the original AIM form or Forms A*/B*. The slightly higher predictive validity of Forms A and B does not seem attributable to sample differences alone. For example, the correlations of the original form Adaptability Composite with 3- and 6-month attrition are similar to those of the alternate forms in each sample—hence, they are higher in the Forms A/B sample (-.14 and -.17, respectively) than in the Forms A*/B* sample (-.12 and -.13, respectively). The similarity continues for 9-, 12-, and 18-month attrition in the Forms A*/B* sample (correlations ranging from -.11 to -.13), but Forms A/B evidence higher correlations (with A*/B*) than does the original form (correlations ranging from -.14 to -.18 as opposed to a correlation of -.14). The higher correlations are attributable to Form A; Form B and the original form perform quite similarly.

The slight advantage in predictive validity estimates for Form A does not arise from a statistical artifact of higher attrition base rates in that sample. Higher attrition rates would mean higher point-biserial correlations were possible in the Form A sample than in the Forms A*/B*. As shown in Table 4.10, however, the opposite situation obtains: The attrition rate is higher in the Forms A*/B* sample than in the Form A sample at all time points.

Table 4.10. Attrition Rates (Percent) at Various Time Points in the Two AIM Alternate Forms Samples

	Attrition (Months)				
Sample	3	6	9	12	18
Forms A/B	7.5	16.2	18.0	19.6	23.3
Forms A*/B*	11.7	16.9	19.1	20.8	25.3

Conclusion

After evaluation of the forms and consideration of the predictive validity analyses, it was determined that both sets of alternate forms would be scaled to the original AIM. The original alternate forms (Forms A and B) performed slightly (but consistently) better than the revised alternate forms (Forms A* and B*) in all three sets of analyses (rank-order correlations, decision consistency, and predictive validity). The differences were slight, however, and there is greater similarity in scale means between the original AIM and Forms A*/B* than between the original AIM and Forms A/B (cf. Heggestad, Young, et al., 1999). The greater discrepancy in scale means for the latter forms would theoretically increase the amount of scaling error. Therefore, both sets of alternate forms (Forms A/B and A*/B*) were scaled to the original AIM. The scaling results appear in the next section.

Before continuing, we wish to note that any change in the scoring of the Adaptability Composite—whether minor (as with the choice not to score one or two item stems) or substantial (as with the adoption of an odds-based decision model as discussed in Chapter 7)—will necessitate recomputation of the decision consistency results. Most changes will necessitate rescaling of the alternate forms to the original form (Levine's and Williams's approach, however, obviates scaling, as all scores would be on the same metric [theta]).

Scaling of AIM Alternate Forms to the Original Form

In an effort to obtain comparable scores, equipercentile scaling was performed for the operational (i.e., original) AIM form and both sets of the AIM alternate test forms (Forms A/B and A*/B*). The scaling task aligned Adaptability Composite scores from the original form and each of the four alternate forms onto a common metric. Although the methods used in equating and scaling are often the same statistically, the purposes of the two procedures are different. Primarily, equating is appropriate when the two test forms have been built to be highly similar in content and statistical properties, and are thought to differ only in terms of difficulty (i.e., strictly parallel forms). Equated scores are considered to be *interchangeable*, because they are adjusted for differences in item difficulty. In contrast, scaling allows one to build score correspondence tables with the explicit acknowledgement that the two forms may not have been built to measure the same content, or at least do not measure the same content to the same degree (i.e., alternate forms) (Mislevy, 1992). Scaled scores are considered to be *comparable* rather than interchangeable due to the nature of differential test construction or measurement issues. The partially ipsative nature of the AIM may have an effect on the degree to which alternate forms measure the same constructs of the operational form and therefore warranted the use of scaling to achieve comparability.

In developing a correspondence table of comparable Adaptability Composite scores for the original form and alternate Forms A/B and A*/B*, we used an analytic equipercentile method (Kolen & Brennan, 1995) to increase efficiency and accuracy over graphical procedures of equipercentile scaling. The equipercentile method identified a function that converted scores from each alternate form to the scale of the original form such that the distribution of converted alternate form scores is equal to the distribution of scores on the operational AIM form in the norm population.

The samples used to effect the scaling were taken from the AIM Grand Research Database (see Chapter 2) and comprised those Regular and Reserve Soldiers who took the original AIM and the alternate Forms A/B or A*/B*. The number of cases having valid composite scores on the initial alternate forms (A/B) were as follows: n = 1,678 for the original AIM and Form B, and n = 1,679 for Form A. The number of cases having valid composite scores on the revised alternate forms (A*/B*) and the original AIM was 6,497. To scale the alternate forms to the original form, each alternate form was directly linked to the original form. Because each respondent completed the original AIM and one set of alternate forms, a single-group design was implemented. The analytic procedures described below allowed the production of a score correspondence table showing each raw score for the Adaptability Composite from Forms A/B and A*/B* and the comparable scores for the Adaptability Composite from the original form.

Equipercentile Scaling Method

The cumulative frequency distribution, $F(x)$, was calculated for the distributions of Adaptability Composite scores on the original form and alternate Forms A/B and A*/B* of the AIM. $F(x)$ represents the proportion of examinees in the population obtaining a score at or below the test score x. Non-integer scores were rounded to the nearest integer and the corresponding frequency was summed with the nearest integer score's frequency. Following Kolen and Brennan (1995, pp. 42–45) and letting K_x represent the number of items on Form X of a test, the percentile rank function, $P(x)$, is

$$\begin{aligned} P(x) &= 100\,\{F(x-1) + [x-(x-0.5)][F(x)-F(x-1)]\}, \quad -0.5 \le x < K_x + 0.5, \\ &= 0, \quad x < -0.5, \\ &= 100, \quad x \ge K_x + 0.5. \end{aligned}$$

The inverse of a percentile rank function, P^{-1}, is used to find a score corresponding to a percentile rank. For a given percentile rank, P^*, the corresponding score is

$$x_U(P^*) = P^{-1}[P^*] = \frac{P^*/100 - F(x_U - 1)}{F(x_U) - F(x_U - 1)} + (x_U - .5), \qquad 0 \leq P^* < 100,$$

$$= K_X + 0.5, \qquad P^* = 100.$$

In the above equation, for $0 \leq P^* < 100$, x_U is the smallest integer score with a cumulative percent that is greater than P^*.

In equipercentile scaling, the objective is to find a score on Form Y that has the same percentile rank as a score on Form X (or vice versa, as the formulas are symmetric). Let $Q(y)$ refer to the cumulative distribution of score y on Form Y. The Form X equipercentile equivalent of score y on Form Y is

$$e_X(y) = P^{-1}[Q(y)]$$

$$= \frac{P(y)/100 - F(x_U - 1)}{F(x_U) - F(x_U - 1)} + (x_U - 0.5).$$

For the current task, equipercentile scaling was performed for the original form and each set of the alternate forms (Forms A/B and A*/B*), creating a table of corresponding Adaptability Composite scores on the original form given an Adaptability Composite score on each of the alternate AIM forms. Corresponding scores for possible score values that were outside the range of observed scores were calculated through interpolation. Smoothing the score distributions and equipercentile relationships did not appear necessary—initial score distributions were smooth and followed a shape likely to be found in the population. The results of the scaling for Forms A/B and A*/B* are presented in Table 4.11 and 4.12, respectively.

Table 4.11. Adaptability Composite Score Correspondence Table for Initial Alternate Forms (A/B) and Original AIM Form.

Alternate AIM Form A Score	Original AIM Score	Alternate AIM Form B Score	Original AIM Score
		91	91
		90	90
		89	88
		88	87
87	91	87	86
86	89	86	84
85	88	85	83
84	86	84	81
83	85	83	80
82	83	82	79
81	82	81	79

Table 4.11. (Continued)

Alternate AIM Form A Score	Original AIM Score	Alternate AIM Form B Score	Original AIM Score
80	80	80	79
79	79	79	78
78	78	78	76
77	75	77	74
76	74	76	73
75	72	75	72
74	71	74	71
73	70	73	70
72	69	72	69
71	67	71	67
70	66	70	66
69	65	69	65
68	64	68	63
67	62	67	62
66	60	66	61
65	59	65	60
64	58	64	59
63	57	63	58
62	56	62	57
61	55	61	56
60	54	60	55
59	53	59	54
58	53	58	53
57	52	57	53
56	51	56	52
55	50	55	51
54	49	54	50
53	48	53	49
52	47	52	48
51	46	51	47
50	46	50	46
49	45	49	45
48	44	48	44
47	43	47	43
46	42	46	42
45	41	45	41
44	40	44	40
43	39	43	40
42	38	42	39
41	37	41	38
40	37	40	37
39	36	39	36
38	35	38	34

Table 4.11. (Continued)

Alternate AIM Form A Score	Original AIM Score	Alternate AIM Form B Score	Original AIM Score
37	34	37	34
36	33	36	33
35	32	35	32
34	32	34	32
33	31	33	31
32	30	32	31
31	30	31	30
30	29	30	29
29	28	29	28
28	28	28	28
27	28	27	26
26	26	26	24
25	25	25	22
24	25	24	22
23	24	23	21
22	22	22	21
21	21	21	19
20	20	20	17
19	20	19	15
18	19	18	14
17	16	17	13
16	15	16	13
15	14	15	12
14	13	14	11
13	12	13	10
12	11	12	10
11	10	11	9
10	9	10	8
9	9	9	7
8	8	8	6
7	7	7	6
6	6	6	5
5	5	5	4
4	4	4	3
3	3	3	3
2	2	2	2
1	1	1	1
0	0	0	0

Note. Maximum possible score on Form A = 87; Form B = 91, original form = 91.

Table 4.12. Adaptability Composite Score Correspondence Table for Revised Alternate Forms (A/B*) and Original AIM Form.*

Alternate AIM Form A* Score	Original AIM Score	Alternate AIM Form B* Score	Original AIM Score
		89	91
		88	90
87	91	87	88
86	89	86	87
85	87	85	86
84	86	84	85
83	84	83	83
82	82	82	82
81	81	81	82
80	79	80	80
79	79	79	79
78	78	78	79
77	77	77	78
76	76	76	78
75	75	75	77
74	74	74	76
73	73	73	75
72	72	72	73
71	71	71	72
70	70	70	71
69	68	69	70
68	67	68	69
67	66	67	68
66	65	66	67
65	64	65	66
64	63	64	65
63	62	63	64
62	61	62	63
61	60	61	62
60	59	60	61
59	58	59	60
58	57	58	59
57	56	57	58
56	56	56	57
55	55	55	56
54	54	54	55
53	53	53	54
52	52	52	53
51	51	51	52
50	50	50	51
49	49	49	50
48	49	48	49

Table 4.12. (Continued)

Alternate AIM Form A* Score	Original AIM Score	Alternate AIM Form B* Score	Original AIM Score
47	47	47	48
46	46	46	47
45	45	45	46
44	44	44	45
43	43	43	44
42	41	42	42
41	40	41	41
40	39	40	40
39	38	39	38
38	37	38	38
37	36	37	37
36	35	36	36
35	34	35	35
34	33	34	34
33	32	33	33
32	32	32	32
31	31	31	31
30	31	30	30
29	30	29	29
28	29	28	28
27	28	27	28
26	27	26	26
25	26	25	26
24	26	24	25
23	25	23	24
22	24	22	23
21	23	21	23
20	22	20	22
19	21	19	20
18	19	18	19
17	19	17	18
16	18	16	17
15	17	15	16
14	16	14	15
13	15	13	14
12	13	12	13
11	12	11	12
10	11	10	11
9	10	9	10
8	9	8	8
7	8	7	7
6	7	6	6
5	6	5	5

Table 4.12. (Continued)

Alternate AIM Form A* Score	Original AIM Score	Alternate AIM Form B* Score	Original AIM Score
4	4	4	4
3	3	3	3
2	2	2	2
1	1	1	1
0	0	0	0

Note. Maximum possible score on Form A* = 87; Form B* = 89, original form = 91.

CHAPTER 5. EFFECTS OF RECRUIT CHARACTERISTICS ON FIRST-TERM ATTRITION

Dan J. Putka and Rodney A. McCloy
Human Resources Research Organization

Premature separation from service among enlisted Soldiers is a complex outcome affected by the interaction of many different factors. For this reason, it is unrealistic to expect a single measure, like AIM, to be highly effective in predicting attrition. We believe an optimal strategy requires the use of several measures which, in combination, provide a broad profile covering a range of individual difference factors (e.g., physical, cognitive, and motivational) that are relevant for predicting first-term attrition. This was the rationale behind the analyses reported in this chapter.

The effort presented here explored how AIM might be used <u>in combination with other measures</u> to enhance attrition prediction. This effort may help us better understand the multiple factors contributing to attrition. However, we now know (from Chapter 3) that AIM findings from research data—such as those used in the analyses below—do not generalize well to the operational setting. For this reason, our more recent model development efforts (which go beyond the time period of this report) used updated versions of the <u>operational</u> database.

This chapter reports on our investigation of the efficacy of the AIM in combination with other potentially operational predictors (e.g., Armed Services Vocational Aptitude Battery [ASVAB] subtests, self-report survey items) for identifying Soldiers who are high risks for attrition during their first term of enlistment. The specific focus of the present investigation was on predicting 18-month attrition status among two groups identified by past research as being at particularly high risk for attrition: Soldiers classified in educational Tier 2, and female Soldiers classified in educational Tier 1 (Sipes, Strickland, Laurence, DiFazio, & Wetzel, 2000). Tier 1 Soldiers have a high school diploma and Tier 2 Soldiers have an alternative education credential.

The primary predictor variables of interest in this investigation can be classified into three groups: (a) AIM scales and the Adaptability Composite, (b) operational predictors (objective measures collected prior to enlistment, such as ASVAB subtest scores, age, years of education, body mass index, and entry pay grade), and (c) portions of the Soldier Reception Survey (SRS)[1] that could be used operationally or for identifying Soldiers in need of subsequent attrition-reducing interventions. The efficacy of these three groups of variables for predicting Soldiers' 18-month attrition status was investigated by conducting four sets of logistic regression analyses.

[1] The Soldier Reception Survey (SRS) is a self-report questionnaire designed to tap the attitudes and experiences of enlisted personnel. The SRS was administered to 70% of all incoming accessions from mid-January through August of 1999 as part of the Army's First Term Project (detailed further in the methods section). Because many items on this survey may be predictive of Soldiers' attrition status (Sipes, Strickland, Laurence, DiFazio, & Wetzel, 2000), SRS items were included in the present investigation.

The first set of logistic regression analyses investigated the ability of the AIM scales to predict 18-month attrition in each of the two groups of interest (i.e., Tier 2 Soldiers and female Tier 1 Soldiers). Currently, three of the seven AIM scales are combined to form the Adaptability Composite. Given the exploratory nature of this investigation, the validity and utility of using these scales separately, as opposed to in their aggregate form (i.e., the Adaptability Composite), was estimated within each group of interest. After fitting these "Adaptability-only" models, the other AIM scales were added to the models to investigate the possibility that they incrementally increase the models' validity and utility.

The second set of logistic regression analyses explored the effects of adding operational predictors to the best combination of AIM variables identified in the first set of analyses. Upon fitting this second set of models, a third set of models was fitted to the data to investigate the validity and utility that might be gained by adding SRS variables to models containing both AIM and operational predictors. The fourth set of models examined how conclusions regarding the salience of measures identified as predictive of attrition in the first three model sets might change if non-operational demographic variables were placed in the model. Specifically, the fourth set of models fitted for the Tier 2 Soldiers examined the effects of including race and gender on the validity of individual predictors identified in previous models as significant predictors of Tier 2 attrition. Similarly, for female Tier 1 Soldiers, the fourth set of models examined the effects that race had on the validity of individual predictors identified in previous models as significant predictors of female Tier 1 attrition. Past research has found these demographics to be important predictors of attrition in the Army (e.g., Sipes et al., 2000); thus evaluating the impact they have on the predictiveness of other variables is an important avenue to investigate. Indeed, without controlling for the effects of these demographic variables, it is difficult to assess whether other predictors identified in the first three sets of models may simply be serving as proxies for race or gender.

Method

Attrition Samples

The 18-month attrition status of each Soldier examined in this investigation was based on the December 2000 update of the AIM Attrition Research Dataset from the Enlisted Master File (EMF). The AIM data examined in the present case were gathered from a large sample of recruits entering the Army between September of 1998 and May 1999.[2] Recruits completed the AIM during in-processing at Army reception battalions. The present investigation used the data from all Regular Army Soldiers who would have had the opportunity to complete at least 18 months of service by the time the December 2000 update of the AIM Attrition Research Dataset occurred. Within this sample, 15,694 Soldiers had valid 18-month attrition data. The overall 18-month attrition rate within this sample was 24.8%. Of the 15,694 Soldiers in this sample, 3,163 were Soldiers who entered the Army in educational Tier 2, and 1,700 were female Soldiers who entered the Army in educational Tier 1. Unless noted otherwise, these latter two subsamples of Soldiers were used for all analyses that follow.

[2] Data collection periods for the SRS and AIM overlapped from mid-January until May of 1999. Thus, SRS data were not available for Soldiers who took the AIM between September of 1998 and early January 1999.

In addition to the AIM Attrition Research Dataset (AIM scores, demographics, and some operational variables), data were also drawn from the First Term Database, which contains a wide array of data collected as part of the Army's First Term Project. The First Term Project is an effort to investigate first-term attrition among FY99 non-prior service enlisted accessions. For the present investigation, data only on Soldiers' SRS scores and other variables not contained in the AIM Attrition Research Dataset (e.g., Soldier height, weight) were drawn from the First Term Database. These data from the First Term Database were merged with data in the AIM Attrition Research Dataset only for those Soldiers who had both AIM data and valid 18-month attrition data. When merging these files, it became apparent that complete sets of operational and SRS data were not available for all individuals who had AIM scores. Thus, the effective sample sizes for constructing 18-month attrition models for both Tier 2 and female Tier 1 Soldiers were reduced. Table 5.1 presents the sample sizes available for each set of analyses, as well as the 18-month attrition rates for Tier 2 and female Tier 1 Soldiers.

Table 5.1 reveals that when operational variables were added to models of attrition containing only AIM variables, the number of valid cases dropped by 4.6% in the Tier 2 sample and by 3.9% in the female Tier 1 sample. When SRS and operational variables were added to models of attrition that contained AIM variables only, the resulting decrease in sample size was 42% in the Tier 2 sample and 48.9% in the female Tier 1 sample. Although this drop in sample size is unfortunate, it is important to note that the attrition rates for both the Tier 2 and female Tier 1 samples remained stable as sample size declined.

Table 5.1. 18-Month Attrition Status for Tier 2 and Female Tier 1 Soldiers

Group/Sample	*n*	18-Month Status		Attrition Rate
		Attritee	Nonattritee	
Tier 2				
AIM	3,163	1,027	2,136	32.5
AIM & Operational	3,019	966	2,053	32.0
AIM, Operational, & SRS	1,835	584	1,251	31.8
Female Tier 1				
AIM	1,700	613	1,087	36.1
AIM & Operational	1,634	586	1,048	35.9
AIM, Operational, & SRS	869	314	555	36.1

Note. For the fourth set of models (i.e., those with non-operational demographics), effective sample sizes were identical to those in the third set of models (i.e., AIM, Operational, and SRS).

Measures

As mentioned in the introduction to this chapter, three sets of variables were examined in the present investigation: AIM variables, operational variables, and variables from the Soldier Reception Survey. The following section details the variables that were examined within each of these sets; the complete listing of variables that were examined in this investigation is presented in Table 5.2.

AIM

In this investigation we examined the predictiveness of all seven AIM scales. Depending on the AIM model being examined, either the AIM Adaptability Composite or its component scales were examined as predictors of attrition.

Operational Predictors

In addition to the AIM variables, the investigation also examined 14 operational variables. Operational variables were defined as those that were objectively assessed prior to enlistment and that could be used as selection tools by the Army. These variables included (a) scores on each of the 10 ASVAB subtests, (b) age at the time of enlistment, (c) entry pay grade, (d) years of education at the time of enlistment, and (e) body mass index at time of enlistment. Although other operational variables were available for investigation, they were excluded from analyses due to their lack of variation in the current samples. Examples include youth program participation prior to enlistment and waiver type (if any) prior to entering the Army.

Table 5.2. Predictor Variables Examined in the Present Investigation

AIM	
AIM Adaptability Composite	AIM Scale D
AIM Scale A	AIM Scale E
AIM Scale B	AIM Scale F
AIM Scale C	AIM Scale G
Operational Predictors	
Age (at accession)	ASVAB: Mechanical Comprehension
ASVAB: Arithmetic Reasoning	ASVAB: Numerical Operations[a]
ASVAB: Automotive Shop	ASVAB: Paragraph Comprehension
ASVAB: Coding Speed[a]	ASVAB: Word Knowledge
ASVAB: Electronics Information	Body Mass Index[b]
ASVAB: General Science	Entry Pay Grade
ASVAB: Mathematical Knowledge	Years of Education at Application
SRS Composite Predictors	
21 SRS Composite Predictors (1-21)	
SRS Single-Item Predictors	
17 SRS Single-Item Predictors (1-17)	
Non-Operational Demographic Predictors	
Gender	
Race / Ethnicity	
White vs. Black, White vs. Hispanic, White vs. Other	

Note. [a] The Coding Speed and Numerical Operations subtests are likely to be dropped from the ASVAB so they may not be readily available for future screening efforts. [b] Body mass index was calculated by dividing a Soldier's weight in kilograms by height in meters squared. Soldiers' heights and weights at enlistment were available from the First Term Database.

Soldier Reception Survey Predictors

A number of variables stemming from the SRS were examined to evaluate the degree to which the SRS might add to the validity and utility of a battery composed solely of AIM and operational variables. The SRS is a 61-question survey that was administered to Army Soldiers within a week of their accession. Two types of questions appear on the SRS: (a) those designed to assess Soldiers' pre-accession cognitions, feelings, and beliefs regarding the Army and their decision to join; and (b) self-report biodata items that ask Soldiers about their past experiences (e.g., "Did you participate in clubs during high school?" "How large was the town you grew up in?"). Many of the SRS questions are multi-part in nature. For example, one question asks Soldiers how important each of 27 reasons was in their decision to join the Army, with each reason rated on a 5-point scale. Thus, although the SRS asked only 61 questions, data on 212 items (potential SRS predictor variables) are available for each Soldier who fully completed the survey.

Given that the intent of this investigation was to explore factors that may supplement the AIM and operational predictors at the time of enlistment, one way of reducing the 212 potential SRS predictor variables was to focus only on those items that may be administered prior to enlistment. In light of the intent of this investigation, several SRS items were eliminated from the pool of potential SRS predictors due to their reliance on respondents' progression through the Delayed Entry Program (DEP) (e.g., how often did you attend DEP activities, likely reasons for leaving within 6 months). Furthermore, SRS biodata items that were objectively assessed elsewhere in Soldiers' files were eliminated (e.g., year of birth, gender, race, highest level of education, component of Army joined). Eliminating these items resulted in 146 potential SRS predictors (items) that could be examined in the present investigation. Because many of the remaining items attempted to tap similar constructs (e.g., physical conditioning, social deviance, confidence, military values, current affective state), and given the desire to reduce redundancy among those items, several steps were taken to form composites based on the items. The steps in forming these composites are briefly described below.

The first step taken to reduce the remaining 146 SRS items was to identify SRS composites used in past research. Eight composites used in past work were identified and used in the present investigation.[3] These 8 composites accounted for 45 of the 146 SRS items, thus reducing total number of single-item SRS predictors to 101.

Reducing the remaining 101 SRS items to a smaller number of composites was achieved via a combination of both empirical (i.e., exploratory factor analysis, empirical keying) and rational methods. Via these methods, 13 new SRS composites were formed (accounting for 65 of the remaining 101 SRS items), leaving 17 single-item SRS predictors.[4] In total, 38 SRS variables (21 composites, 17 single-items) were examined in the present investigation. A final list of the SRS predictors are listed in Table 5.2. Inter-item reliabilities for the SRS composites ranges from .50 to .94, though only 5 of the composites had reliabilities of less than .70.

[3] Due to the sensitivity surrounding the content of these composites, the citation for the work from which these composites were drawn is not provided.

[4] In forming the new composites, we dropped seven SRS items that were initially considered for inclusion in a given composite because they substantially decreased the reliability of the given composite when included and had low item-composite correlations. The dropped items are not included among the 17 remaining single-item SRS predictors that were examined in this investigation.

Modeling Attrition within Each Sample

Several steps were taken to evaluate the relationship between each of the predictors identified in Table 5.2 and Soldiers' 18-month attrition status within the two samples. First, zero-order point-biserial correlations were computed between each continuously-scaled predictor and attrition status. For dichotomous categorical predictors, such as marital status, phi coefficients were computed to assess their degree of relationship to attrition status. For polytomous categorical predictors, a set of dichotomous comparisons among the categories of the predictor was created to examine the relationships between Soldiers' membership in one referent category (compared to each of the other categories) and attrition. For example, three sets of dichotomous predictors were created for race; one dichotomy compared White and Black examinees, another compared White and Hispanic examinees, and a final dichotomy compared White with all other races (not Black or Hispanic). Separate phi coefficients were then generated to index each race dichotomy-attrition status relationship.

In addition to point-biserial correlations and phi coefficients, a series of hierarchical logistic regression analyses was used to model attrition in each sample (Tier 2, and Tier 1 females) as a function of the predictors examined. Given that many of the predictor variables examined in this research were scaled differently, all predictors were standardized across the overall 18-month attrition sample prior to fitting all models examined. Standardizing the predictors facilitated interpretation of their corresponding conditional odds ratios from each logistic regression analysis.

The conditional odds ratio for a predictor is formed by raising the mathematical quantity *e* to the power of the standardized partial beta weight of that predictor in a given logistic regression model. In the present analyses, conditional odds ratios reflect the change in odds of a Soldier attriting, given a one-unit increase on the predictor variable of interest, holding all other predictors constant. Because all continuous-scale predictors were standardized before they entered the logistic regression analyses, a one-unit change in these predictors reflects a change of one standard deviation. For categorical variables, a one-unit change represents a change in category (e.g., from single to married). Conditional odds ratios greater than 1.00 indicate an increase in the likelihood of attrition associated with a one-unit increase on the predictor, whereas those less than 1.00 indicate a decrease in the likelihood of attrition associated with a one-unit increase on the predictor of interest (holding the levels of all other predictors constant). Thus, a conditional odds ratio of 1.10 for a predictor would indicate that for every 1 unit increase on the predictor of interest, a Soldier is 1.1 times as likely to attrit compared to the Soldier at the next lowest unit on that predictor, holding all other predictors constant.

In the sections that follow, a description of each logistic regression model examined is provided. Details on how predictors were eliminated from these models using a modified backward stepwise elimination procedure that started with the full set of variables examined in each set of analyses conducted (Sets 2 through 4 only) are also provided. Following this discussion, methods for comparing the relative validity and utility of the various models constructed are discussed.

Set 1: AIM Only

The first set of models fitted within each sample examined the efficacy of various combinations of the AIM scales as predictors of Soldiers' 18-month attrition status. Four specific models were examined. First, a model was fitted that included only the Adaptability Composite (AIM Model 1). Second, a model that included only the Adaptability components (i.e., Scales A through C) was fitted to the data (AIM Model 2). After that, a third model was fitted to the data that included the Adaptability Composite as well as the other AIM scales (i.e., Scales D through G) (AIM Model 3). Lastly, a fourth model was fitted to the data that included all seven AIM scales separately (AIM Model 4). The purpose of fitting this first set of models was to identify the best combination of AIM scales to retain for examination in subsequent models. Subsequent models investigated whether adding operational and/or SRS predictors significantly improved the validity and utility of a model of attrition consisting solely of AIM-based predictors.

Set 2: AIM + Operational Predictors

The second set of models fitted to the data within each sample examined the efficacy of the AIM in combination with the operational variables as predictors of 18-month attrition status. Fitting the "AIM + Operational" models to the data was done using two separate modeling strategies. The first step in both the first and second strategies involved fitting a logistic regression model that consisted of the best combination of AIM predictors identified in the first set of models (AIM Models 1, 2, 3, or 4) along with all operational predictors. Upon fitting this first full model to the data, the two modeling strategies diverged. The first analysis strategy eliminated only those operational variables that failed to meet "importance" criteria (detailed in later sections). These importance criteria were essentially a function of three quantities: (a) the magnitude of the predictor's zero-order point-biserial correlation with attrition, (b) the magnitude of its partial logistic regression beta weight, and (c) the significance of its partial logistic regression weight. All AIM variables in this first strategy were retained, regardless of their status on the importance criteria. Upon eliminating the operational variables that failed to meet the importance criteria, a subsequent reduced hierarchical logistic regression model was fitted to the data that included the AIM variables and the reduced set of operational variables. This process of fitting a model and eliminating the operational variables that failed to meet the importance criteria was repeated until all operational variables that remained in the model met the importance criteria. Unlike the first strategy, the second analysis strategy involved taking the results of fitting the full model to the data and subsequently eliminating all predictors (AIM and operational variables) that did not meet the importance criteria. Subsequent reduced models were estimated until all variables included in the models met the importance criteria.

The purpose of adopting two modeling strategies to identify salient variables for predicting 18-month attrition in the present investigation was twofold. Given that the primary interest of this investigation was to assess whether any other variables would add to the validity and utility of the AIM as a predictor of attrition, deleting AIM variables from any given model would not allow one to make that comparison. Therefore, the first analysis strategy provides an incremental approach to building a composite of potential predictors of attrition. Specifically, under this first strategy, which will subsequently be referred to as the "incremental-fit strategy," factors identified as important predictors of attrition at earlier stages in the model building process (e.g., in Set 1) remain in subsequent models of attrition regardless of whether they meet importance criteria in reductions of these later models (e.g., the AIM in the case of Set 2, or the AIM and operational variables in the case of Set 3 described below).

Although such a practical approach to modeling may be justified given the purpose of the present investigation, it is not wholly satisfying from a statistical perspective. Specifically, if a variable no longer accounts for a significant portion of unique variance when other variables are added to the model, and if importance criteria indicate that the variable adds little validity or utility to one's model, then it should be dropped from the model to achieve greater parsimony. Thus, the second modeling strategy, which will subsequently be referred to as the "best-composite strategy," adopts an approach to model fitting that removes *any* predictor from a model if it does not meet the established importance criteria. The goal of this second modeling strategy is simply to find the most parsimonious combination of predictors that results in levels of validity and utility comparable to those found in the full model (i.e., prior to eliminating variables based on importance criteria).

The benefit of adopting these two modeling strategies within the current investigation is that one can compare the resulting models formed by the two strategies to see if (a) any appreciable loss in the validity or utility of the more practically oriented incremental-fit-based models occurs compared to their corresponding best-composite models, and (b) any AIM variables would drop out of subsequent models as other variables are added (e.g., in the best-composite models).

Set 3: AIM + Operational Predictors + SRS Predictors

The third set of models fitted within each sample examined the gains in validity and utility observed when the SRS predictors were added to a model containing the AIM and operational predictors. As was the case with the second set of models, two strategies were used to fit these models to the data. In this case, however, the starting points (or full models) for each strategy differed. For the "incremental-fit" model, all SRS predictors were added to a model containing the AIM and operational predictors identified by the reduced incremental-fit model from Set 2. This full model was fitted to the data and subsequently reduced by eliminating only SRS predictors that failed to meet importance criteria, much as operational predictors were eliminated for the incremental-fit model in Set 2.

For the "best-composite" model, all operational and SRS predictors were added to a model containing the best combination of AIM variables identified in Set 1. This full model was fitted to the data, and subsequently reduced by eliminating any predictor (including AIM and operational variables) that did not meet the importance criteria. The reduction processes for the incremental-fit model and best-composite models were repeated until all eligible variables within each model met the established importance criteria.

Set 4: Demographics + AIM + Operational Predictors + SRS Predictors

The purpose of exploring the fourth set of models was to determine whether any predictors identified by the Set 3 models as potentially important predictors of 18-month attrition lost their importance as a result of including the non-operational demographic variables as a first step in the model. As was the case with the models examined in Sets 2 and 3, both the incremental-fit and best-composite strategies were used to fit models to the data. One difference between the modeling effort in Set 4 relative to the other sets was that the incremental-fit strategy did not allow for the elimination of any variables, as it simply added the incremental-fit model from Set 3 to the non-operational demographics available for analysis (no reduction was conducted, since all variables were identified as important in previous analyses). On the other hand, for the best-composite strategy, all AIM variables identified in Set 1, all operational predictors, and all SRS were added to a model containing only the non-operational demographics. This model was subsequently fitted to the data, and any variables that failed to meet the importance criteria described below were eliminated from the model.

Eliminating Individual Predictors from Full Models of Attrition

Because as many as 61 predictors were available for comparison in the present investigation (depending on the model examined), one of the primary decisions to be made in modeling Soldiers' 18-month attrition status was how to produce parsimonious models of attrition. One of the dangers of traditional stepwise logistic regression strategies for model building is that variables that are important predictors may be eliminated (e.g., backward stepwise elimination) or fail to be added to a model (e.g., forward stepwise selection), due to problems arising from multicollinearity among the predictors (Agresti, 1996). For the present investigation, a modified backward stepwise elimination procedure was employed to help alleviate some of the problems cited above.

The present procedure differed from traditional backward stepwise elimination in three ways. First, the modified procedure did not remove predictors one at a time; rather, it removed predictors in groups. Second, the modified procedure did not eliminate predictors based on which one had the lowest *p*-value associated with its partial regression weight relative to other predictors in the equation. Third, the modified procedure required that predictors fail to meet two criteria before they would be eliminated from subsequent models (rather than just a single criterion). The first criterion was based on a predictor's adjusted zero-order relationship with 18-month attrition status (i.e., point-biserial correlation for continuous predictors, phi coefficient for categorical predictors). The adjustment to a predictor's zero-order coefficient was based on the strength of its relationship with attrition status, holding all other predictors in the equation constant (i.e., its logistic regression partial beta weight). The formula used to make this adjustment [i.e., $r_{xy\ adj} = r_{xy} * (b_{xy} * 10)$] is a variation on Hoffman's (1962) index of the relative weight (importance) of predictors in the context of multiple regression.

The adjusted correlation coefficient for each predictor was then transformed to z using Fisher's (1925) transformation. The predictor was retained if its z-transformed adjusted coefficient differed from 0 by 1.5 or more standard deviations. If the predictor's z-transformed coefficient failed to meet this criterion, it was not eliminated from the model unless it also failed to meet the second criterion. The second criterion was its p-value from the Wald's test of its partial regression weight. If the p-value for the Wald's test for the predictor was less than 0.1, the predictor was retained. Thus, for a predictor to be eliminated from a model, the predictor would need to have a z-transformed correlation less than 1.5 standard deviations away from 0 and have a p-value greater than or equal to 0.1 (with regard to the significance of its partial regression weight). The effects of using this modified elimination strategy (in terms of creating reduced models of attrition) was that variables that would have been eliminated based on traditional criteria alone (e.g., lack of significance in p-values for partial betas) were retained if they had sizable zero-order coefficients yet still predicted (at least to a small degree) attrition in the presence of the other predictors. Given the exploratory nature of this investigation, the number of predictors involved, and that an ancillary goal of these analyses was to safely identify variables that were clearly *not* important predictors of attrition (particularly in the presence of other available variables), adopting a more conservative criterion for *excluding* variables in the present investigation appeared warranted.

Evaluating the Validity and Utility of Different Models

Several comparisons were made between the models examined in the present investigation with regard to both their validity and utility for predicting Soldiers' attrition status. These comparisons can be broken down into two basic types: those that focused on the validity of the models for predicting 18-month attrition status, and those that focused on examining the utility of the models for identifying recruits who would likely attrit by 18 months.[5] We compared the relative fit of nested models by examining the difference between maximized log-likelihood ratios of the reduced and full models and testing the significance of that difference against a χ^2 distribution with degrees of freedom equal to the difference in the number of predictors in each model. In all analyses, Soldiers' attrition status was coded as 1 (attrit) or 0 (nonattrit).

Based on the results of the logistic regression analyses, several indexes were used to assess the validity of each model. First, we generated a point-biserial correlation between the predicted probability of a Soldier's attrition based on the model in question and actual attrition status. We also calculated Cohen's d (effect size)[6] index for each model. Cohen's d reflects the mean standardized difference between attrits and nonattrits in terms of their predicted probabilities of attrition as specified by a given model. In addition to this index, Negelkerke's R^2 was calculated for each model. Negelkerke's R^2 indicates the proportion of variance in Soldiers' likelihood of attrition that can be accounted for by predictors in the model of interest (Negelkerke, 1991).[7]

In addition, to the above indexes of model validity, several indexes from Signal Detection Theory were employed to evaluate the validity of each model. First, ROC (Receiver Operating Characteristic) curves were generated. In the context of examining attrition, ROC curves display the tradeoff between the hit rate (i.e., the proportion of attriting Soldiers who tested positive for attrition) and the false positive rate (i.e., the proportion of nonattriting Soldiers who tested positive for attrition) for each possible cut score for a given "test for attrition."[8] Each point on a

---

[5] The type of "utility" examined in the present investigation refers to the expected quality of the decisions made for a given cut score on the predictor composites examined by each model. We did not investigate the dollar-cost utility associated with a given cut score.

[6] The point-biserial correlation and Cohen's d provide the same type of information regarding the predictive efficacy of a model. Both statistics are presented however, for ease of comparison to past and future work.

[7] Negelkerke's R^2 was used as opposed to a more traditional Cox-Snell R^2 because Negelkerke's has the desirable property of ranging from 0 to 1 (unlike Cox-Snell). This makes its interpretation more similar to the R^2 reported in traditional multiple-regression.

[8] In the context of this investigation, coding attrition status as 1 (attrit) and 0 (nonattrit) resulted in predicted probabilities of attrition that correlated negatively with the AIM. Because many models consisted of multiple AIM scales, a Soldier's predicted probability of attriting served as the Soldier's score on the "test for attrition." This specification facilitates the use of the traditional medical Signal Detection Theory (SDT) model, where attrition can be viewed as the "disease," and the composite of predictors used in a given model can be viewed as a test to detect the disease (i.e., a test for attrition). In light of this observation, we discuss cut scores on the "test for attrition" (i.e., the predicted probabilities resulting from a particular logistic model) in terms of percentages of Soldiers falling at or above a certain probability of attriting. To clarify, setting higher cut scores on the test of attrition would correspond

ROC curve corresponds to a specific cut score on the diagnostic test (here the test for attrition). Each cut score has an associated hit rate (a y-coordinate) and false positive rate (an x-coordinate) (see Figure 3.1 in Chapter 3 for a sample ROC curve). Following the trace of a ROC curve from left to right reveals the tradeoff that *increasing the cut score* on a given test for attrition would have on these two proportions.

One particularly useful piece of information resulting from the generation of ROC curves is the area under the ROC curve (*AUC*), which here provides an indicator of the accuracy with which a given model predicts a Soldier's attrition status (Hanley & McNeil, 1982). Specifically, the *AUC* index for a given model reflects the expected proportion of times that an attritee would score higher than a nonattritee on the given attrition composite if a pair of Soldiers (one attritee and one nonattritee) were repeatedly selected at random from each group. For example, an *AUC* value of .70 means that attritees would be expected to score higher than nonattritees on the given attrition composite 70% of the time. *AUC* values typically range between .50 (indicating equal probability of attritees and nonattritees scoring higher) and 1.0 (indicating that attritees would always score higher than nonattritees). An additional benefit of *AUC* values is that they have a theoretical sampling distribution, which allows confidence intervals to be calculated, thus providing an inferential means to compare the accuracy of two or more models.[9]

To examine the utility with which each model successfully identified Soldiers who were more likely to attrit, additional indices based on Signal Detection Theory (i.e., predictive values of positive tests and false acceptance rates) were generated for a variety of potential percentage-based cut scores for each model. Specifically, the predictive value of a positive test (i.e., PVPT—the proportion of Soldiers who attrited, given that they tested positive for attrition based on a particular cut score) and false acceptance rate (i.e., the proportion of Soldiers who attrited, given they tested negative for attrition based on a particular cut score) were calculated for each model under a variety of potential cut scores. In this investigation, one can also interpret the false acceptance rate as the expected 18-month attrition rate should the examined cut score be implemented. The present investigation examined three cut scores that were targeted to evaluate the utility of the models if approximately 10%, 30%, or 50% of applicants were screened out on the basis of their composite scores of the predictors comprising each model. Examining a variety of potential cut scores allows investigators to evaluate the utility of models at a variety of different points on a given attrition composite's scale.

Summary

Given the numerous models that were fitted as part of this investigation for both Tier 2 and female Tier 1 Soldiers, the results of the model-fitting efforts are divided into two main sections. The first presents results for Tier 2 Soldiers only, while the second presents results for

---

to requiring more stringent requirements for diagnosing a Soldier as an attrition risk, effectively screening out *fewer* Soldiers. Specifically, screening out the top 10% of scorers on a test for attrition would mean you are screening out the 10% of Soldiers you believe are most likely to attrit.

[9] Inferential comparisons between *AUC* values of different models in this investigation are conservative (i.e., estimated confidence intervals are wider than what they would likely be in reality) because the method used to generate standard errors assumes that the models being compared were estimated on independent samples of data (Hanley & McNeil, 1982). In this investigation, we compared *AUC* values for alternative AIM models using a single sample, thus violating the independence assumption.

female Tier 1 Soldiers. Within each section, only reduced models resulting from the two modeling strategies introduced earlier (i.e., the incremental-fit and best-composite strategies) are presented.

Tier 2 Sample Results

Descriptives

The means and standard deviations for Tier 2 Soldiers on all continuously scaled predictor variables are presented in Table 5.3. To provide a contrast for this information, means and standard deviations for Tier 1 Soldiers (both male and female) are also presented in this table, along with Cohen's *d* values (contrasting Tier 2 and Tier 1 Soldiers) for each of the predictors. The predictors in Table 5.3 are divided into four sets: AIM, operational predictors, SRS composite predictors, and SRS single-item predictors. Within each set of predictors, specific variables are ranked in order from highest to lowest with regard to the absolute value of their observed *d* value. Thus, predictors at the top of each set are ones on which Tier 2 and Tier 1 Soldiers differed most, while predictors near the bottom are those on which these groups of Soldiers differed least. Positive *d* values indicate that Tier 2 Soldiers scored higher than Tier 1 Soldiers on the predictor of interest. Both the number and percentage of Soldiers falling within each category of the demographic predictor variables for Tier 2 and Tier 1 Soldiers are presented in Table 5.4.

Table 5.3. Comparison of Tier 2 and Tier 1 Soldiers on the Continuously-Scaled Predictor Variables

Predictor	Tier 2 Soldiers			Tier 1 Soldiers			
	n	*M*	*SD*	*n*	*M*	*SD*	*d*
AIM							
AIM Scale B	3,167	22.77	5.01	12,228	24.59	4.80	-0.377
AIM Adaptability Composite	3,166	53.89	11.08	12,204	56.28	10.14	-0.231
AIM Scale C	3,173	9.42	2.96	12,249	9.86	2.89	-0.153
AIM Scale D	3,173	18.75	3.75	12,247	19.05	3.56	-0.083
AIM Scale G	3,170	1.22	1.35	12,248	1.26	1.39	-0.031
AIM Scale A	3,171	21.71	5.38	12,232	21.82	5.03	-0.022
AIM Scale F	3,172	21.65	4.71	12,240	21.58	4.52	0.016
AIM Scale E	3,171	17.25	4.04	12,246	17.29	3.91	-0.010
Operational Predictors							
Years of Education at Application	3,074	10.87	0.57	11,829	12.27	1.04	-1.466
Entry Pay Grade	3,151	1.05	0.23	12,179	1.48	0.84	-0.568
ASVAB: Mathematical Knowledge	3,148	50.41	5.91	12,171	53.62	7.51	-0.444
ASVAB: Automotive Shop	3,148	52.34	7.78	12,171	49.86	8.29	0.303
ASVAB: Coding Speed	3,148	51.48	6.86	12,171	53.64	7.34	-0.298
ASVAB: Numerical Operations	3,148	51.82	7.39	12,171	53.76	7.24	-0.266
Age	3,151	20.19	2.72	12,179	20.85	3.15	-0.217

Table 5.3. (continued)

Predictor	Tier 2 Soldiers			Tier 1 Soldiers			
	n	*M*	*SD*	*n*	*M*	*SD*	*d*
Operational Predictors– Continued							
ASVAB: Mechanical Comprehension	3,148	54.16	7.66	12,171	52.34	8.55	0.216
ASVAB: Paragraph Comprehension	3,148	53.85	4.98	12,171	52.66	6.19	0.199
ASVAB: Word Knowledge	3,148	53.53	4.41	12,171	52.45	5.68	0.199
ASVAB: Electronics Information	3,148	51.57	6.66	12,171	50.10	7.67	0.197
Body Mass Index	3,151	24.17	3.58	12,179	24.82	3.77	-0.175
ASVAB: General Science	3,148	52.39	6.36	12,171	51.76	7.84	0.083
ASVAB: Arithmetic Reasoning	3,148	52.49	6.25	12,171	52.08	7.53	0.057
SRS Composite Predictors							
SRS Composite: 11	2,446	3.54	3.07	8,028	2.22	2.45	0.506
SRS Composite: 9	2,446	-0.21	0.65	8,028	0.06	0.67	-0.405
SRS Composite: 3	2,443	4.22	0.81	8,021	4.10	0.86	0.131
SRS Composite: 6	2,446	0.25	0.53	8,028	0.30	0.58	-0.093
SRS Composite: 7	2,446	3.38	0.89	8,028	3.30	0.87	0.087
SRS Composite: 2	2,442	3.55	0.92	8,022	3.48	0.90	0.080
SRS Composite: 17	2,439	3.09	0.87	7,999	3.03	0.86	0.078
SRS Composite: 20	2,302	1.32	1.01	7,576	1.39	1.01	-0.074
SRS Composite: 8	2,444	6.06	1.01	8,011	5.99	1.03	0.063
SRS Composite: 5	2,441	3.19	0.93	8,024	3.24	0.90	-0.048
SRS Composite: 19	2,437	2.23	0.99	7,981	2.19	1.00	0.046
SRS Composite: 1	2,438	0.84	0.31	7,983	0.83	0.32	0.041
SRS Composite: 10	2,443	2.85	1.08	8,019	2.81	1.09	0.036
SRS Composite: 12	2,423	0.91	1.08	7,958	0.95	1.07	-0.031
SRS Composite: 21	2,302	0.45	0.62	7,576	0.43	0.63	0.031
SRS Composite: 16	2,437	1.64	0.87	7,982	1.66	0.88	-0.020
SRS Composite: 4	2,443	4.22	0.86	8,022	4.21	0.83	0.015
SRS Composite: 18	2,439	3.60	1.12	7,995	3.59	1.12	0.012
SRS Composite: 13	2,439	3.69	1.01	8,004	3.70	1.03	-0.011
SRS Composite: 14	2,439	3.39	0.92	8,000	3.39	0.96	0.007
SRS Composite: 15	2,439	4.09	0.87	8,002	4.09	0.90	-0.003
SRS Single-Item Predictors							
SRS Single-Item: 13	2,444	3.46	0.70	8,020	3.90	0.69	-0.651
SRS Single-Item: 5	2,396	2.33	0.76	7,972	2.66	0.66	-0.477
SRS Single-Item: 16	2,430	3.57	1.49	7,990	3.32	1.46	0.175
SRS Single-Item: 15	2,437	3.35	0.98	7,994	3.21	0.89	0.147
SRS Single-Item: 2	3,151	0.39	0.84	12,179	0.29	0.73	0.138
SRS Single-Item: 10	2,443	3.73	1.08	8,009	3.69	1.04	0.030
SRS Single-Item: 14	2,440	0.36	0.48	8,005	0.35	0.48	0.010
SRS Single-Item: 3	2,442	3.28	1.00	8,015	3.27	0.99	0.008

Table 5.4. Comparison of Tier 2 and Tier 1 Soldiers on the Demographic Predictor Variables

Predictor	Tier 2 Soldiers		Tier 1 Soldiers	
	n	%	*n*	%
Gender				
Male	2,877	91.3	10,444	85.8
Female	274	8.7	1,735	14.2
Race / Ethnicity				
White	2,410	76.5	7,368	60.5
Black	398	12.6	2,838	23.3
Hispanic	249	7.9	1,318	10.8
Other	94	3.0	655	5.4

Zero-order correlations between each of the predictor variables and Tier 2 Soldiers' 18-month attrition status are presented in Table 5.5. In addition to these values, mean predictor scores of Tier 2 attrits and nonattrits are presented, along with their corresponding Cohen's *d* values. As was the case with Table 5.3, predictors are presented by set and sorted in descending order based on the absolute value of their observed *d* value. Positive *d* values indicate that nonattrits scored higher than attrits.

Based on Table 5.5, the AIM predictors with the largest difference between those Tier 2 Soldiers who attrited, and those who did not, are the Adaptability Composite ($d = 0.220$) and Scale D ($d = 0.213$) with the attrition group mean falling below the nonattrition group mean for both variables. Minimal differences were apparent between the attrits and nonattrits on Scales E and G, with differences of less than 0.02 standard deviations between the group means. These latter findings are consistent with past research that has investigated relationships between the AIM scales and Soldiers' attrition status at 3, 6, and 9 months (Heggestad, Young, Strickland, & Rumsey, 1999).

Only 2 of the 14 operational variables exhibited mean differences between attrits and nonattrits that were greater than or equal to one-tenth of a standard deviation: ASVAB Mechanical Comprehension ($d = 0.119$) and ASVAB Electronics Information ($d = 0.100$). Of the 14 operational variables, 8 revealed less then one-twentieth of a standard deviation difference between the mean of attrits and the mean of nonattrits. Such findings indicate minimal differences exist between 18-month attrits and nonattrits on the majority of the operational variables examined in the present investigation.

The SRS predictors with the largest difference between Tier 2 attrits and nonattrits were the SRS Composites 4 and 5, with the attritee mean falling 0.301 and 0.252 standard deviations below the nonattritee mean, respectively.

Table 5.5. Zero-Order Correlations between Predictors and 18-Month Attrition Status for Tier 2 Soldiers

Variable	*n*	*r*	$M_{Attrits}$	$M_{Nonattrits}$	*d*
AIM					
AIM Adaptability Composite	3,163	-.103	52.26	54.69	0.220
AIM Scale D	3,170	-.099	18.22	19.01	0.213
AIM Scale A	3,168	-.093	20.99	22.06	0.199
AIM Scale C	3,170	-.092	9.02	9.60	0.198
AIM Scale F	3,169	-.075	21.14	21.89	0.160
AIM Scale B	3,164	-.073	22.24	23.02	0.156
AIM Scale G	3,167	.008	1.23	1.21	-0.018
AIM Scale E	3,168	.002	17.26	17.24	-0.004
Operational Predictors					
ASVAB: Mechanical Comprehension	3,145	-.056	53.53	54.45	0.119
ASVAB: Electronics Information	3,145	-.047	51.11	51.78	0.100
ASVAB: Automotive Shop	3,145	-.044	51.84	52.57	0.094
Age	3,148	-.042	20.02	20.27	0.091
ASVAB: Word Knowledge	3,145	.042	53.79	53.40	-0.089
ASVAB: Arithmetic Reasoning	3,145	-.027	52.25	52.60	0.057
ASVAB: Mathematical Knowledge	3,145	-.022	50.22	50.50	0.047
ASVAB: Numerical Operations	3,145	-.015	51.66	51.90	0.033
Years of Education at Application	3,071	-.015	10.85	10.87	0.031
Body Mass Index	3,148	.011	24.23	24.14	-0.023
ASVAB: Coding Speed	3,145	.010	51.58	51.42	-0.022
ASVAB: General Science	3,145	.007	52.44	52.35	-0.015
Entry Pay Grade	3,148	.001	1.05	1.05	-0.002
ASVAB: Paragraph Comprehension	3,145	.001	53.85	53.84	-0.001
SRS Composite Predictors					
SRS Composite: 4	2,441	-.140	4.05	4.30	0.301
SRS Composite: 5	2,439	-.117	3.04	3.27	0.252
SRS Composite: 2	2,440	-.088	3.44	3.61	0.189
SRS Composite: 1	2,436	-.078	0.81	0.86	0.167
SRS Composite: 6	2,444	.075	0.31	0.22	-0.161
SRS Composite: 9	2,444	-.074	-0.28	-0.17	0.158
SRS Composite: 8	2,442	-.068	5.96	6.11	0.146
SRS Composite: 11	2,444	.067	3.83	3.40	-0.143
SRS Composite: 3	2,441	-.059	4.15	4.25	0.127
SRS Composite: 7	2,444	-.058	3.30	3.41	0.124
SRS Composite: 21	2,300	.051	0.49	0.43	-0.109
SRS Composite: 14	2,437	-.050	3.33	3.42	0.106
SRS Composite: 20	2,300	-.049	1.25	1.35	0.104
SRS Composite: 18	2,437	-.043	3.53	3.64	0.091

Table 5.5. (continued)

Variable	n	r	$M_{Attrits}$	$M_{Nonattrits}$	d
SRS Composite Predictors—Continued					
SRS Composite: 15	2,437	-.042	4.03	4.11	0.090
SRS Composite: 12	2,421	.040	0.97	0.88	-0.085
SRS Composite: 10	2,441	-.037	2.79	2.87	0.078
SRS Composite: 17	2,437	-.027	3.06	3.11	0.058
SRS Composite: 13	2,437	.007	3.70	3.68	-0.016
SRS Composite: 19	2,435	.006	2.24	2.23	-0.013
SRS Composite: 16	2,435	-.004	1.64	1.65	0.010
SRS Single-Item Predictors					
SRS Single-Item: 16	2,428	-.082	3.40	3.66	0.176
SRS Single-Item: 6					
A vs. B	765	-.066	.	.	
A vs. C	1,784	-.054	.	.	
SRS Single-Item: 8					
A vs. B	2,034	.058	.	.	
A vs. C	1,456	.055	.	.	
SRS Single-Item: 17	2,421	-.050	.	.	
SRS Single-Item: 12	2,429	-.045			
SRS Single-Item: 5	2,394	.045	2.38	2.31	-0.097
SRS Single-Item: 15	2,435	-.045	3.28	3.38	0.096
SRS Single-Item: 13	2,442	.036	3.49	3.44	-0.078
SRS Single-Item: 9	2,436	.036	.	.	
SRS Single-Item: 4	2,406	-.032	.	.	
SRS Single-Item: 14	2,438	-.028	0.34	0.36	0.059
SRS Single-Item: 11	2,441	-.025	3.69	3.74	0.054
SRS Single-Item: 7	2,410	-.021	.	.	
SRS Single-Item: 1	3,148	.019	.	.	
SRS Single-Item: 3	2,440	-.013	3.26	3.29	0.027
SRS Single-Item: 10	2,432	.010	.	.	
SRS Single-Item: 2	3,148	.002	0.40	0.39	-0.005
Non-Operational Demographic Predictors					
Gender[a]	3,148	.143	.	.	
Race / Ethnicity[b]					
White vs. Black	2,805	-.034	.	.	
White vs. Hispanic	2,656	-.064	.	.	
White vs. Other	2,501	.009	.	.	

Note. Soldiers' 18-month attrition status was coded as 0 (nonattrit) or 1 (attrit). r values reported for categorical variables are phi coefficients. Means, standard deviations, and d statistics were not computed for categorical variables because mean values reflect proportions rather than mean scale scores (see Table 5.4). [a] Responses to the "gender" item were coded as 0 (male) or 1 (female). [b] Responses to the "race/ethnicity" item were coded as 0 (White) or 1 (Black, Hispanic, or Other depending on the analysis conducted).

Model Set 1: Evaluating Different AIM Combinations

The first set of models fitted to the data for predicting 18-month attrition status among Tier 2 Soldiers evaluated various combinations of the AIM scales. Table 5.6 presents the results of fitting these models to the Tier 2 data.

Table 5.6. Comparison of Different AIM Combinations for Predicting Tier 2 Soldiers' 18-Month Attrition Status

Model / Statistics	AIM Model			
	1	2	3	4
Conditional Odds Ratios[a]				
AIM Adaptability Composite	0.815***	.	0.864**	.
AIM Scale A	.	0.893*	.	0.917
AIM Scale B	.	0.971	.	0.990
AIM Scale C	.	0.884**	.	0.906*
AIM Scale D	.	.	0.893*	0.886**
AIM Scale E	.	.	1.107*	1.088
AIM Scale G	.	.	1.086*	1.092*
AIM Scale F	.	.	0.922	0.940
Model-Level Validity				
R	0.105	0.111	0.131	0.134
d	-0.224	-0.238	-0.281	-0.289
Nagelkerke R^2	0.015	0.016	0.022	0.024
AUC	0.558	0.562	0.573	0.575
AUC 95% C.I.	(0.537, 0.580)	(0.540, 0.583)	(0.552, 0.595)	(0.554, 0.597)
Model-Level Utility				
Top 10 % Cut				
PVPT	45.3	46.9	47.5	49.7
False Acceptance Rate	31.1	30.8	30.8	30.5
Top 30 % Cut				
PVPT	37.7	40.6	40.7	40.7
False Acceptance Rate	30.2	29.0	28.9	28.9
Top 50 % Cut				
PVPT	36.0	36.1	36.9	36.8
False Acceptance Rate	29.0	28.8	27.9	28.1

Note. PVPT refers to the predictive value of a positive test (i.e., the proportion of Soldiers who attrited, given that they tested positive for attrition based on the given cut). [a] Conditional odds ratios (OR) greater than 1.0 reflect an increase in the likelihood of attrition as scores on the predictor increase. OR less than 1.0 reflect a decrease in the likelihood of attrition as scores on the predictor increase (holding all other predictors in the model constant). Model 1 contains only the Adaptability Composite (n = 3,163). Model 2 contains each of the Adaptability Composite scales (n = 3,163). Model 3 contains the Adaptability Composite, as well as the four non-Adaptability AIM scales (n = 3,158). Model 4 contains all AIM scales (n = 3,158). * $p < .05$, ** $p < .01$, *** $p < .001$.

Adaptability Composite vs. Adaptability Scales

The first comparison made was between models examining the predictiveness of the Adaptability Composite by itself versus the AIM scales that constitute it. Because these models are not nested, their relative fit to the data could not be compared using differences between their maximized log-likelihood ratios. The other criteria described in the methods section were used to compare the relative quality of these models to one another.

Table 5.6 reveals little difference in either the validity or utility of these two models for predicting 18-month attrition status among Tier 2 Soldiers (Model 1: Adaptability Composite, Model 2: Adaptability scales). All indices used to examine model-level validity and utility revealed great similarity in the models. For example, the point-biserial correlation for the Adaptability Composite model was .105, whereas the correlation for the Adaptability scales model was .111. Comparing *AUC* values provides an inferential test of quality of the two models for predicting attrition. The *AUC* value for the model with the Adaptability scales only (0.562) did not fall outside of the 95% confidence interval surrounding the *AUC* value for the model with the Adaptability Composite only (0.558), thus suggesting little or no validity would be gained by looking at the Adaptability scales separately instead of the Adaptability Composite. Based on the *AUC* value for the Adaptability Composite model, attrits would be expected to receive lower Adaptability scores than nonattrits 56.2% of the time (using the current Adaptability Composite scoring). Such values indicate that the Adaptability Composite alone is not distinguishing very well between attrits and nonattrits, as an *AUC* value of 0.50 indicates pure chance.

Table 5.6 also presents information regarding the utility of the Adaptability Composite by itself as a predictor of 18-month attrition status among Tier 2 Soldiers. Predictive values of positive tests (PVPT) and false acceptance rates (expected attrition rate if the cut score were implemented) are also presented for a variety of potential cut scores. Depending on the cut score used, the percentage of Soldiers who attrited, given that they were identified as being a likely attrit by the Adaptability-Composite-only model fell between 36% (cutting out the top 50%) and 45.3% (cutting out the top 10%). Examining the false acceptance rates associated with the various cut scores reveals the expected base rate of attrition for Soldiers in Tier 2 should that cut be adopted operationally. Based on the results presented in Table 5.6, the expected rate of 18-month attrition among Tier 2 Soldiers if the Adaptability Composite were adopted would be between 29% (cutting out the top 50%) and 31.1% (cutting out the top 10%). Given that the base rate of 18-month attrition among Tier 2 Soldiers in this sample is 32.5%, implementing the cut scores examined here can be expected to reduce attrition among Tier 2 Soldiers between 4.3% (cutting out the top 10%) and 10.8% (cutting out the top 50%).

Adaptability Composite + Other AIM Scales

Given the similarity of the two "Adaptability only" models reviewed above, both for the sake of parsimony and because the Adaptability Composite is what is being used operationally in the GED Plus program, subsequent comparisons focused on adding to that model. A model with all AIM scales entered separately was also fitted to the data for purposes of comparison.

A model with Adaptability and the other AIM scales (i.e., Model 3 in Table 5.6) significantly increased model fit over the model containing only the Adaptability Composite (χ^2

(4) = 17.8, $p < .01$). The point-biserial correlation between the predictor composite and 18-month attrition status increased from .105 to .131. Although the *AUC* value for the present model (0.573) was higher than the *AUC* value for the Adaptability-Composite-only model (0.558), the difference was not statistically significant, as the former fell within the latter's 95% confidence interval. Furthermore, the PVPT values and false acceptance rates did not appreciably differ between the two models. Specifically, the PVPT values for the present model were only between 0.9 (cutting out the top 50%) and 3.0 (cutting out the top 30%) percentage points higher than those for the Adaptability-Composite-only model. Moreover, the expected rates of attrition (false acceptance rates) were only between 0.3 (cutting out the top 10%) and 1.3 (cutting out the top 30%) percentage points higher than those for the Adaptability-Composite-only model.

As was the case with the two Adaptability-only models, there were minimal differences between Models 3 (Adaptability Composite and other AIM scales) and 4 (all AIM scales). Although small differences existed between the four AIM models examined, adding the non-Adaptability scales appeared to improve model fit slightly (based on significant differences between log-likelihood ratios). Interestingly, two scales were not significant predictors of Tier 2 Soldiers' attrition status (Scale B, Models 2 and 4; Scale F, Models 3 and 4). Nevertheless, given the significant relationship between these two scales and Tier 2 Soldiers' attrition status (based on zero-order correlations), they were retained for analysis in subsequent models. Given the results of analyzing the four AIM models presented above, and the minimal differences observed between Models 3 and 4, subsequent model sets examined the impact of adding other predictors to the AIM Model 3 on predicting Tier 2 Soldiers' 18-month attrition status.

Model Sets 2–4: Adding Extra Predictors to AIM Model 3

Tables 5.7 and 5.8 present the results of adding operational and SRS predictors to the third AIM model identified above. The format of Table 5.7 is similar to the latter half of Table 5.6 and presents information regarding the validity and utility of each set of models examined. The predictor variables listed in Table 5.8 are broken down by type (e.g., AIM, operational, SRS) for each set of models examined. Specific predictor variables within each type are sorted in descending order of their estimated conditional odds ratios.

Only those predictor variables that were retained in at least one of the models are listed in Table 5.8. Predictor variables that did not meet inclusion criteria for any of the models are unlikely to be important predictors of Tier 2 Soldiers' attrition status.

Set 2: AIM + Operational Predictors

In the second set of models, a hierarchical logistic regression model was constructed where AIM Model 3 was entered at the first step and all operational predictors were entered in at the second step. Upon fitting this full model, 8 of 14 predictors were eliminated based on the inclusion criteria described earlier. All AIM variables were retained. Thus, within Set 2, the incremental-fit and best-composite models comprised the same predictors.

Table 5.7. Model-Level Comparison of Different Models of Tier 2 Soldiers' 18-Month Attrition Status

Model / Statistics	AIM Model 3	Set 2: AIM + Operational	Set 3: AIM + Operational + SRS		Set 4: All Predictors	
			Incremental	Best	Incremental	Best
Model-Level Validity						
R	0.131	0.165	0.274	0.270	0.298	0.298
d	-0.281	-0.357	-0.609	-0.601	-0.667	-0.669
Nagelkerke R^2	0.022	0.037	0.099	0.096	0.117	0.117
AUC	0.573	0.587	0.651	0.659	0.664	0.676
AUC 95% C.I.	(0.552, 0.595)	(0.559, 0.615)	(0.624, 0.678)	(0.632, 0.686)	(0.637, 0.691)	(0.649, 0.702)
Model-Level Utility						
Top 10 % Cut						
PVPT	47.5	47.4	64.6	60.1	64.1	61.3
False Acceptance Rate	30.8	30.8	28.8	28.9	28.9	28.7
Top 30 % Cut						
PVPT	40.7	42.3	47.7	48.1	49.4	50.0
False Acceptance Rate	28.9	28.3	25.9	25.1	25.1	24.2
Top 50 % Cut						
PVPT	36.9	38.8	41.8	41.7	43.1	43.3
False Acceptance Rate	27.9	26.1	23.0	22.2	21.7	20.5

Note. AIM Model 3 contains the Adaptability Composite, as well as the four non-Adaptability AIM scales (n = 3,158). Set 2 contains AIM and operational predictors (n = 3,125). There is only one Set 2 model listed because the models fitted using the incremental-fit and best-composite strategies resulted in identical reduced sets of predictors. Set 3 contains AIM, operational, and SRS predictors ($n_{\text{Incremental}}$ = 1,950, n_{Best} = 1,907). Set 4 contains all predictors (non-operational demographics included) ($n_{\text{Incremental}}$ = 1,950, n_{Best} = 1,905). "Incremental" indicates models fitted using the incremental-fit strategy. "Best" indicates models fitted using the best-composite strategy.

The reduced model and full model (i.e., AIM Model 3 and all operational predictors) provided a similar degree of fit to the data, even though the former contained eight fewer operational variables, $\Delta\chi^2(8) = 2.19$, *n.s.* As an additional check on whether the variables that were eliminated from the full model added anything to the reduced model, all excluded variables were correlated with the standardized residuals resulting from the reduced model to see if any significant relationships emerged (suggesting a variable may incrementally increase validity if put back into the reduced model). No eliminated variable was significantly related to the residuals of the reduced model. The variables retained in the reduced model for Set 2, as well as the conditional odds ratios associated with each of them, appear in Table 5.8 under the Set 2 heading.

Based on this Set 2 model, individuals who score highly on the AIM Adaptability Composite (*OR* = .879)[10], AIM Scale D (*OR* = .880), ASVAB Mechanical Comprehension test (*OR* = .888), or ASVAB Electronics Information test (*OR* = .884) are less likely to attrit than Soldiers who score low on those scales. Furthermore, Soldiers who achieve high scores on the ASVAB Word Knowledge test (*OR* = 1.270), AIM Scale E (*OR* = 1.099), or AIM Scale G (*OR* = 1.095) are more likely to attrit than Soldiers who score low on these scales.

[10] "*OR*" refers to the conditional odds ratio of a predictor.

Table 5.8. Predictor Level Comparison of Different Models of Tier 2 Soldiers' 18-Month Attrition Status

Predictor	Conditional Odds Ratios				
	Set 2: AIM + Operational	Set 3: AIM + Operational + SRS		Set 4: All Predictors	
		Incremental	Best	Incremental	Best
AIM					
AIM Scale G	1.095	1.132	1.152	1.145	1.171
AIM Scale E	1.099	1.119		1.103	
AIM Adaptability Composite	0.879	0.972		1.000	
AIM Scale F	0.928	0.969		0.940	0.960
AIM Scale D	0.880	0.872	0.846	0.874	0.866
Operational Predictors					
ASVAB: Word Knowledge	1.270	1.166	1.155	1.124	1.123
Body Mass Index			1.137		1.182
ASVAB: Automotive Shop					1.067
ASVAB: Coding Speed	1.080	0.983		0.980	
ASVAB: Numerical Operations	0.932	0.987		0.950	
Age	0.928	0.966		0.959	0.878
ASVAB: Mechanical Comprehension	0.888	0.863	0.869	0.888	0.877
ASVAB: Electronics Information	0.884	0.848	0.831	0.902	0.873
Years of Education at Application			0.795		0.782
SRS Composite Predictors					
SRS Composite: 17		1.150	1.125	1.149	1.134
SRS Composite: 21		1.130	1.130	1.144	1.140
SRS Composite: 13		1.125	1.145	1.106	1.136
SRS Composite: 11		1.097	1.075	1.129	1.092
SRS Composite: 6		1.101	1.077	1.107	1.075
SRS Composite: 8			1.064		1.068
SRS Composite: 2		1.076	1.057	1.060	1.038
SRS Composite: 14			0.942		0.920
SRS Composite: 7		0.901	0.909	0.909	0.917
SRS Composite: 20		0.892	0.896	0.902	0.900
SRS Composite: 16		0.880	0.892	0.893	0.903
SRS Composite: 9		0.867	0.874	0.906	0.914
SRS Composite: 5		0.889	0.877	0.896	0.890
SRS Composite: 4		0.821	0.828	0.851	0.858

Table 5.8. (continued)

Predictor	Conditional Odds Ratios				
	Set 2: AIM + Operational	Set 3: AIM + Operational + SRS		Set 4: All Predictors	
		Incremental	Best	Incremental	Best
SRS Single-Item Predictors					
SRS Single-Item: 8					
A vs. B		1.371	1.422	1.290	1.335
A vs. C		1.425	1.516	1.360	1.446
SRS Single-Item: 5		1.112	1.118	1.087	1.094
SRS Single-Item: 9		1.086		1.087	
SRS Single-Item: 14		0.888		0.890	
SRS Single-Item: 16		0.895	0.896	0.880	0.877
SRS Single-Item: 4		0.773	0.786	0.774	0.792
SRS Single-Item: 6					
A vs. B		0.646	0.661	0.639	0.659
A vs. C		0.682	0.683	0.673	0.679
Non-Operational Demographic Predictors					
Gender				2.425	2.525
Race / Ethnicity					
White vs. Black				0.768	0.846
White vs. Hispanic				0.824	0.769
White vs. Other				1.395	1.219

Note. The coding of all categorical demographic variables presented in this table can be found in Table 5.5. Set 2 contains AIM and operational predictors. Set 3 contains AIM, operational, and SRS predictors. Set 4 contains all predictors (non-operational demographics included). "Incremental" indicates models fitted using the incremental-fit strategy. "Best" indicates models fitted using the best-composite strategy.

In terms of the validity and utility of this model relative to AIM Model 3, there appears to be a small but statistically significant improvement, at least in terms of log-likelihood criteria. Specifically, when the reduced set of operational variables was added to the model containing only the Adaptability Composite and other AIM scales (i.e., AIM Model 3), there was a statistically significant improvement in fit, ($\Delta\chi 2(6) = 32.88$, $p < .001$). In terms of point-biserial correlations with attrition status, the composite based on the present model resulted in a slightly higher correlation ($r = .165$) relative to AIM Model 3 ($r = .134$). In terms of Signal Detection Theory criteria, the AUC for the current model (0.587) was greater than the AUC for AIM Model 3 (0.573), yet this difference was not statistically significant.

In terms of the quality of decisions made based on the current model compared to the AIM-only model, no improvement was apparent. Focusing only on the PVPT values and false acceptance rate if one were to cut out the top 10% of scorers, the current model resulted in only a 0.1 percentage point decrease in PVPT values compared to AIM Model 3, and an identical false acceptance rate (i.e., 30.8%). Thus, if the Army were to (a) implement either the current model or AIM Model 3 and (b) screen out Soldiers who score in the top 10% of these composites, the expected rate of attrition for Tier 2 Soldiers would be the same (30.8%) in both cases, which represents a 4.3% improvement on the base rate of 18-month attrition among Tier 2 recruits.

Set 3: AIM + Operational Predictors + SRS Predictors

When the SRS variables were considered in terms of whether they added any validity to models containing AIM and operational predictors, the two strategies of model fitting discussed earlier were followed. First, a model was fitted to the data that simply added all the SRS predictors to the reduced model from Set 2 above. This model was subsequently reduced to eliminate any SRS predictors that did not meet the criteria for inclusion described earlier (incremental-fit model). In parallel to this analysis, another model was fitted to the data, this one containing AIM Model 3, all operational predictors, and all SRS predictors. This model was reduced to eliminate any predictor (including AIM and operational predictors) that did not meet the criteria for inclusion described earlier (best-composite model).

In the process of reducing the incremental-fit and best-composite models several SRS predictors were eliminated (19 for the incremental-fit model 19 for the best-composite model, with much overlap). Both the reduced incremental-fit Set 3 model ($\Delta\chi^2(19) = 8.26$, *n.s.*) and the reduced best-composite Set 3 model ($\Delta\chi^2(31) = 16.93$, *n.s.*) provided similar degrees of fit to the data compared to their respective full models, despite the incremental-fit model containing 19 fewer predictors than its full model, and the best-composite model containing 31 fewer predictors than its full model (3 AIM predictors, 9 operational predictors, and 19 SRS predictors were eliminated). As a check to see whether any of the excluded variables could potentially add anything to their respective reduced models, the eliminated variables were correlated with the standardized residuals of the full models from which they were dropped. All resulting correlations were nonsignificant, thus indicating the variables likely would not incrementally increase the validity of either reduced model. Retained variables in the reduced models of Set 3 and the conditional odds ratios associated with each of them are presented in Table 5.8 under the Set 3 heading.

In fitting these two models to the data, areas of convergence in terms of which predictors appeared to be most important for predicting attrition status were identified. Based on both the incremental-fit and best-composite models, two SRS single-item predictors were most related to attrition. Specifically, Soldiers who scored in category B ($OR_{\text{Incremental}} = 1.371$; $OR_{\text{Best}} = 1.422$), or category C ($OR_{\text{Incremental}} = 1.425$; $OR_{\text{Best}} = 1.516$) on SRS Single-Item Predictor 8 were more likely to attrit compared to Soldiers who scored in category C. Conversely, Soldiers who scored in category B ($OR_{\text{Incremental}} = .646$; $OR_{\text{Best}} = .661$) or category C ($OR_{\text{Incremental}} = .682$; $OR_{\text{Best}} = .683$) on SRS Single-Item Predictor 6 were less likely to attrit than Soldiers who scored in category A.

One of the primary reasons for fitting the best-composite model in Set 3 was to see whether any of the AIM or operational variables retained in the Set 2 reduced model would be eliminated when SRS variables were added to the model. Comparing the Set 2 and Set 3 best-composite models revealed the following results. First, although the AIM Adaptability Composite, AIM Scale E, and AIM Scale F were all retained in the Set 2 model, these three variables were eliminated when the SRS variables were added in Set 3. Similarly, ASVAB Coding Speed, Numerical Operations, and age were all retained in the Set 2 model, yet were eliminated from the best-composite model in Set 3. In each case, these results suggest that the variance in attrition being tapped by these Set 2 variables is being accounted for by the SRS variables.

In terms of the validity and utility of the Set 3 incremental-fit and best-composite models relative to models containing just the AIM and operational variables, adding the SRS variables appeared to have a substantial impact. Specifically, when the reduced set of SRS variables was added to the AIM and operational variables in the Set 2 model, there was a significant improvement in the fit of the model to the data ($\Delta\chi^2(21) = 89.75$, $p < .001$). That is, the Set 3 incremental-fit model provided a better fit to the data than the Set 2 model. Similarly, adding the reduced set of SRS variables to the reduced set of AIM and operational variables in the third step of the Set 3 best-composite model significantly improved model fit ($\Delta\chi^2(21) = 89.27, p < .001$). Compared to the 18-month attrition point-biserial correlation for the Set 2 model ($r = .165$), both Set 3 models exhibited substantial increases in validity ($r = .274$, incremental-fit model; $r = .270$, best-composite model). The *AUC* values for the two Set 3 models were also significantly greater than the *AUC* value for the Set 2 model. Specifically, both the *AUC* value for the Set 3 incremental-fit model (0.651) and the *AUC* value for the Set 3 best-composite model (0.659) exceeded the upper bound of the 95% confidence interval surrounding the Set 2 model's *AUC* value (0.587). The *AUC* values for the Set 3 models suggest that if pairs of Tier 2 Soldiers were randomly selected from those who attrited and those who did not, the Soldier from the attrition group would score higher on a composite formed from the Set 3 models approximately 65% of the time.

In terms of the quality of decisions based on Set 3 models compared to the Set 2 model, improvement was again apparent. Focusing only on PVPT values when cutting out the top 10% of Tier 2 Soldiers (as that most nearly reflects what may be used in practice), the Set 3 models resulted in 12.7 (best-composite model) and 17.2 (incremental-fit model) percentage point increases compared to the Set 2 model. Also apparent was the impact that adding the SRS variables had on the expected Tier 2 attrition rate. Namely, including the SRS predictors identified by the Set 3 models and cutting out the top 10% of Tier 2 Soldiers based on the resulting composite scores would be expected to yield an 11.4% (incremental-fit model false acceptance rate = 28.8%) or 11.1% (best-composite model false acceptance rate = 28.9%) decrease in attrition relative to the Tier 2 base rate of 32.5%. Note that this is compared to an improvement on the base rate of attrition of only 4.3% when using a top 10% cut if only the AIM and operational variables are used (Set 2 model).

Set 4: Demographics + AIM + Operational Predictors + SRS Predictors

The final set of Tier 2 models examined the impact that entering demographics had on the results obtained from the predictors retained by the models in Set 3. Once again two strategies were followed. One strategy was simply to enter the demographics at the first step and then enter the reduced incremental-fit model from Set 3.[11] The other strategy was to enter the demographic variables first and then enter all other predictors, eliminating variables from the model using the importance criteria established earlier (Set 4 best-composite model). In reducing the latter model, 28 variables were eliminated (2 AIM predictors, 7 operational predictors, and

[11] The incremental-fit model for Set 4 was not reduced as no variables identified from the Set 3 incremental-fit model were subject to elimination from the model.

19 SRS predictors). Even in eliminating the 28 predictors, Set 4's reduced best-composite model provided a similar degree of fit to the data relative to a model with all predictor variables entered ($\Delta\chi^2(28) = 14.57$, *n.s.*). As a check to see whether any of the eliminated variables could potentially add anything to the reduced model, the eliminated variables were correlated with the standardized residuals of the reduced model from which they were dropped. Once again, all correlations were nonsignificant, thus indicating the eliminated variables were not likely to add incrementally to the validity of the reduced model. The variables retained for the Set 4 best-composite model and those examined for the incremental-fit model, as well as the conditional odds ratios associated with each of them, are presented in Table 5.8 under the Set 4 heading.

In fitting both Set 4 models to the data, areas of convergence in terms of what predictors appeared to be most important for predicting attrition status were identified. For both the incremental-fit model and the best-composite model, the most salient predictor of attrition was gender ($OR_{\text{Incremental}} = 2.420$; $OR_{\text{Best}} = 2.525$). These conditional odds ratios suggest that Tier 2 females were about 2.5 times as likely to attrit as their Tier 2 male counterparts, holding all other variables constant. This finding is quite consistent with past research and is part of the reason that female Tier 1 Soldiers were examined in the second sample of this investigation. Based on both the incremental-fit and best-composite models, the two next-strongest predictors of attrition once again dealt with Soldiers' scores on SRS Single-Item Predictors 6 and 8. Specifically, individuals who scored in category B ($OR_{\text{Incremental}} = 1.290$, $OR_{\text{Best}} = 1.335$), or C ($OR_{\text{Incremental}} = 1.360$; $OR_{\text{Best}} = 1.446$) on 8, were more likely to attrit compared to Soldiers who scored in category A. Conversely, Soldiers who scored in category B ($OR_{\text{Incremental}} = .639$; $OR_{\text{Best}} = .659$) or C ($OR_{\text{Incremental}} = .673$; $OR_{\text{Best}} = .679$) on 6, were less likely to attrit than Soldiers who scored in category A.

One of the primary reasons for examining the Set 4 models was to see if any predictors retained in the Set 3 best-composite model would be eliminated if demographics were first considered in the model. No predictors retained in the Set 3 best-composite model were excluded from the best composite model in Set 4, but some variables that were retained in the Set 4 best-composite model were not retained in the Set 3 best-composite model (AIM Scale F, ASVAB Automotive Shop subtest, and age). These results suggest that taking into account the demographics had little impact on the contribution of the other predictors that were already included in the model from earlier stages of analysis. Given that the demographics added to the models in Set 4 cannot be used operationally to select Soldiers, the impact that they had on the overall validity and utility relative to the other sets of models will not be detailed here. Nevertheless, note that the validity and utility of the Set 4 models appears to be very similar to their Set 3 counterparts (particularly based on a top 10% cut for utility purposes), suggesting that inclusion of the demographics would not likely make a large difference on the utility of an overall composite of attrition predictors.

Tier 2 Attrition Model Summary

Figure 5.1 provides a graphical summary of the utility of the Tier 2 attrition models examined in the present investigation. Specifically, Figure 5.1 displays the percentage of Tier 2 recruits who attrited within each decile of the predictor composite based on the reduced models for each set of analyses performed. Given the similarity of the best-composite and incremental-fit models within each set of analyses, only the best-composite models were plotted in Figure 5.1.

The results of the model fitting process above suggests that the Army may have much to gain by further exploring the SRS variables identified here as potentially salient predictors of Tier 2 Soldiers' 18-month attrition status. While the AIM and current operational variables were able to predict modest amounts of variation in Soldiers' attrition status, SRS variables increased both the validity and utility of the resulting model. Specifically, the expected decrease in attrition achieved by cutting out the top 10% of Tier 2 Soldiers based on the composite comprising AIM and operational variables was estimated to be about 4%, whereas adding SRS variables to the same composite yielded an expected drop of 11%.

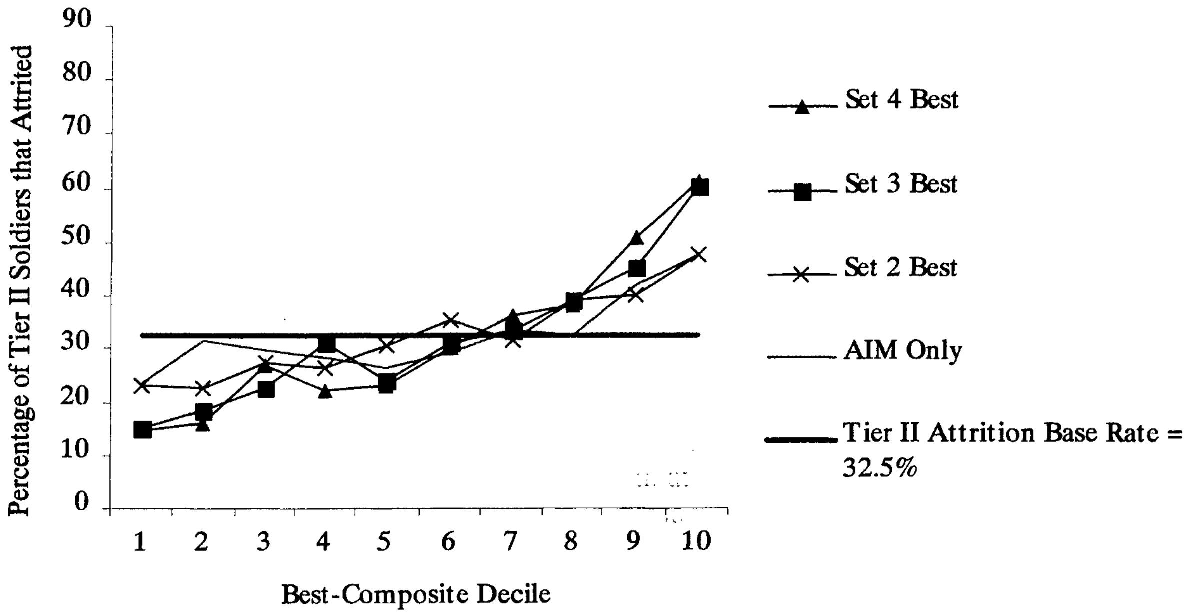


Figure 5.1. Observed percentage of Tier 2 Soldiers that attrited at or before 18 months by best-composite model decile.

Another finding was that the AIM Adaptability Composite and AIM Scale E failed to account for any significant variation in Tier 2 Soldiers' 18-month attrition status once SRS variables were included in the model. Nevertheless, both remained significant predictors of attrition when only operational variables were added to a model containing the AIM predictors. Given that not all SRS variables identified in this investigation as salient predictors of attrition may be used operationally, future research should be conducted to test the possibility that the aforementioned AIM variables retain their significance if only SRS variables that can be used operationally are used in the model.

Consistent with past research, gender emerged as the strongest predictor of attrition among Tier 2 Soldiers, with females being about 2.5 times as likely to attrit as males. After gender, several SRS predictors, notably SRS Single-Item Predictors 6 and 8 were salient for predicting attrition. Such findings suggest that the SRS may provide fruitful content for future work seeking to identify factors that are predictive of first-term Soldier attrition among Tier 2 Soldiers.

In sum, future research should be conducted to (a) determine if the variables identified in the present study as potentially salient predictors of 18-month attrition status cross-validate well in other samples of Tier 2 recruits, (b) establish a rationale for why the variables identified are predictive of attrition, and (c) identify which of the salient SRS predictors might be used operationally during the enlistment process to help identify Tier 2 recruits at greatest risk of attrition.

Female Tier 1 Sample Results

Descriptives

The means and standard deviations for female Tier 1 Soldiers on all continuously scaled predictor variables are presented in Table 5.9. To provide a contrast for this information, means and standard deviations for male Tier 1 Soldiers are also presented in this table, along with Cohen's *d*-values (contrasting female and male Tier 1 Soldiers) for each of the predictors. The predictors in Table 5.9 are divided into four sets and, within each set, specific predictor variables are rank ordered from highest to lowest with regard to the absolute value of their observed *d* value. Thus, predictors at the top are ones on which female and male Tier 1 Soldiers differed most, while predictors near the bottom are those on which these sets of Soldiers differed least. Positive *d* values indicate that female Tier 1 Soldiers scored higher than male Tier 1 Soldiers on the predictor of interest. Both the number and percentage of Soldiers falling within each category of the categorical demographic variables for female and male Tier 1 Soldiers are presented in Table 5.10.

Table 5.9. Comparison of Female and Male Tier 1 Soldiers on the Continuously-Scaled Predictor Variables

Predictor	Female Tier 1 Soldiers			Male Tier 1 Soldiers			
	n	*M*	*SD*	*n*	*M*	*SD*	*d*
AIM							
AIM Scale F	1,720	22.82	4.38	10,397	21.38	4.51	0.321
AIM Scale B	1,721	25.87	4.34	10,385	24.38	4.84	0.311
AIM Scale E	1,724	18.28	4.18	10,399	17.12	3.84	0.298
AIM Scale C	1,721	9.21	2.94	10,405	9.97	2.87	-0.261
AIM Scale G	1,722	1.09	1.28	10,404	1.29	1.41	-0.142
AIM Scale D	1,722	19.36	3.43	10,402	19.00	3.57	0.100
AIM Scale A	1,720	21.43	5.13	10,391	21.89	5.01	-0.091
AIM Adaptability Composite	1,714	56.50	9.73	10,369	56.25	10.21	0.025
Operational Predictors							
ASVAB: Automotive Shop	1,733	43.47	5.89	10,438	50.92	8.16	-0.947
ASVAB: Mechanical Comprehension	1,733	47.54	7.51	10,438	53.14	8.45	-0.674
ASVAB: Electronics Information	1,733	46.03	6.71	10,438	50.77	7.61	-0.634
ASVAB: Coding Speed	1,733	57.04	7.05	10,438	53.07	7.24	0.550
Body Mass Index	1,735	23.56	2.99	10,444	25.03	3.84	-0.395
ASVAB: Numerical Operations	1,733	55.86	6.20	10,438	53.41	7.34	0.340
Entry Pay Grade	1,735	1.68	0.98	10,444	1.44	0.81	0.285
ASVAB: Mathematical Knowledge	1,733	55.28	6.84	10,438	53.34	7.58	0.259
Years of Education at Application	1,669	12.50	1.25	10,042	12.24	0.99	0.255

Table 5.9. (continued)

Predictor	Female Tier 1 Soldiers			Male Tier 1 Soldiers			
	n	*M*	*SD*	*n*	*M*	*SD*	*d*
Operational Predictors – Continued							
ASVAB: Paragraph Comprehension	1,733	53.91	5.33	10,438	52.46	6.30	0.237
ASVAB: Word Knowledge	1,733	53.36	5.10	10,438	52.30	5.76	0.187
Age	1,735	21.32	3.72	10,444	20.78	3.04	0.174
ASVAB: General Science	1,733	50.92	7.18	10,438	51.90	7.93	-0.124
ASVAB: Arithmetic Reasoning	1,733	51.50	7.08	10,438	52.18	7.59	-0.090
SRS Composite Predictors							
SRS Composite: 11	1,170	1.42	1.96	6,858	2.36	2.49	-0.390
SRS Composite: 9	1,170	-0.07	0.65	6,858	0.09	0.67	-0.233
SRS Composite: 15	1,162	4.24	0.85	6,840	4.06	0.91	0.198
SRS Composite: 10	1,169	2.99	1.11	6,850	2.78	1.08	0.193
SRS Composite: 4	1,170	4.08	0.89	6,852	4.23	0.81	-0.188
SRS Composite: 6	1,170	0.21	0.48	6,858	0.32	0.60	-0.181
SRS Composite: 16	1,163	1.53	0.80	6,819	1.68	0.89	-0.173
SRS Composite: 18	1,161	3.73	1.16	6,834	3.56	1.11	0.150
SRS Composite: 19	1,163	2.06	1.00	6,818	2.21	1.00	-0.149
SRS Composite: 13	1,164	3.83	1.02	6,840	3.68	1.03	0.145
SRS Composite: 20	1,096	0.85	0.86	6,480	0.96	0.90	-0.126
SRS Composite: 17	1,164	2.95	0.85	6,835	3.04	0.87	-0.111
SRS Composite: 1	1,165	0.86	0.31	6,818	0.82	0.33	0.100
SRS Composite: 8	1,166	6.08	1.01	6,845	5.98	1.03	0.096
SRS Composite: 1	1,170	3.43	0.92	6,852	3.49	0.89	-0.068
SRS Composite: 5	1,169	3.28	0.91	6,855	3.23	0.90	0.052
SRS Composite: 12	1,155	0.92	1.10	6,803	0.95	1.06	-0.026
SRS Composite: 14	1,163	3.37	1.01	6,837	3.39	0.96	-0.024
SRS Composite: 7	1,170	3.29	0.89	6,858	3.30	0.87	-0.018
SRS Composite: 3	1,169	4.11	0.88	6,852	4.10	0.86	0.013
SRS Composite: 21	1,096	0.78	0.78	6,480	0.77	0.84	0.008
SRS Single-Item Predictors							
SRS Single-Item: 5	1,162	2.90	0.63	6,810	2.61	0.65	0.445
SRS Single-Item: 13	1,169	4.14	0.64	6,851	3.86	0.69	0.405
SRS Single-Item: 14	1,166	0.25	0.43	6,839	0.37	0.48	-0.256
SRS Single-Item: 2	1,735	0.44	0.90	10,444	0.26	0.69	0.244
SRS Single-Item: 15	1,165	3.07	0.71	6,829	3.24	0.91	-0.191
SRS Single-Item: 11	1,163	3.86	1.08	6,846	3.67	1.03	0.186
SRS Single-Item: 16	1,165	3.44	1.47	6,825	3.30	1.46	0.098
SRS Single-Item: 3	1,166	3.30	1.01	6,849	3.26	0.99	0.036

Table 5.10. Comparison of Female and Male Tier 1 Soldiers on the Categorical Predictor Variables

Predictor	Female Tier 1		Male Tier 1	
	n	%	*n*	%
Race / Ethnicity				
White	936	53.9	6,432	61.6
Black	541	31.2	2,297	22.0
Hispanic	158	9.1	1,160	11.1
Other	100	5.8	555	5.3
Marital Status Upon Entry				
Single	1,347	77.6	9,057	86.7
Married	388	22.4	1,387	13.3

Zero-order point-biserial correlations between each of the predictor variables and female Tier 1 Soldiers' 18-month attrition status are presented in Table 5.11. In addition to these values, mean predictor scores of female Tier 1 attrits and nonattrits are presented along with their corresponding Cohen's *d* values. As was the case with Table 5.9, predictors are presented by set and sorted in descending order based on the absolute value of their observed Cohen's *d*-value. Positive *d* values indicate that nonattrits scored higher than attrits on the predictor of interest.

Table 5.11. Zero-Order Correlations between Predictors and 18-Month Attrition Status for Female Tier 1 Soldiers

Predictor	*n*	*r*	M_{Attrits}	$M_{\text{Nonattrits}}$	*d*
AIM					
AIM Scale C	1,712	-.149	8.63	9.54	0.314
AIM Adaptability Composite	1,705	-.143	54.64	57.52	0.300
AIM Scale A	1,711	-.124	20.59	21.91	0.260
AIM Scale F	1,711	-.096	22.24	23.12	0.202
AIM Scale B	1,712	-.073	25.43	26.09	0.153
AIM Scale D	1,713	-.046	19.15	19.48	0.096
AIM Scale E	1,715	.019	18.37	18.21	-0.039
AIM Scale G	1,713	-.010	1.08	1.10	0.022
Operational Predictors					
ASVAB: Automotive Shop	1,724	.139	44.54	42.83	-0.292
ASVAB: Electronics Information	1,724	.074	46.64	45.62	-0.153
ASVAB: Coding Speed	1,724	-.067	56.39	57.38	0.141
ASVAB: Word Knowledge	1,724	.065	53.77	53.08	-0.136
ASVAB: Mathematical Knowledge	1,724	-.064	54.69	55.59	0.132
ASVAB: General Science	1,724	.059	51.44	50.56	-0.124
ASVAB: Numerical Operations	1,724	-.058	55.37	56.12	0.121
Years of Education at Application	1,660	-.042	12.41	12.52	0.088
Body Mass Index	1,726	.042	23.72	23.46	-0.087

Table 5.11. (continued)

Predictor	n	r	$M_{Attrits}$	$M_{Nonattrits}$	d
Operational Predictors – Continued					
Entry Pay Grade	1,726	-.035	1.63	1.70	0.073
ASVAB: Mechanical Comprehension	1,724	.034	47.85	47.32	-0.070
ASVAB: Paragraph Comprehension	1,724	-.027	53.70	54.00	0.056
Age	1,726	.009	21.35	21.27	-0.020
ASVAB: Arithmetic Reasoning	1,724	.009	51.54	51.41	-0.018
SRS Composite Predictors					
SRS Composite: 5	1,163	-.201	3.03	3.41	0.431
SRS Composite: 4	1,164	-.195	3.84	4.20	0.416
SRS Composite: 9	1,164	-.143	-0.20	-0.01	0.302
SRS Composite: 2	1,164	-.126	3.27	3.51	0.266
SRS Composite: 8	1,160	-.125	5.91	6.17	0.263
SRS Composite: 11	1,164	.114	1.73	1.26	-0.240
SRS Composite: 18	1,155	-.113	3.56	3.84	0.237
SRS Composite: 7	1,164	-.089	3.18	3.35	0.187
SRS Composite: 14	1,157	-.078	3.27	3.43	0.163
SRS Composite: 21	1,090	.073	0.86	0.74	-0.153
SRS Composite: 1	1,159	-.071	0.83	0.87	0.149
SRS Composite: 15	1,156	-.061	4.17	4.28	0.127
SRS Composite: 20	1,090	-.056	0.79	0.89	0.117
SRS Composite: 17	1,158	-.049	2.89	2.98	0.103
SRS Composite: 3	1,163	-.046	4.06	4.15	0.097
SRS Composite: 6	1,164	.045	0.24	0.20	-0.095
SRS Composite: 12	1,149	.026	0.96	0.90	-0.055
SRS Composite: 19	1,157	-.024	2.03	2.08	0.050
SRS Composite: 13	1,158	.019	3.86	3.82	-0.040
SRS Composite: 10	1,163	.009	3.00	2.98	-0.018
SRS Composite: 16	1,157	-.004	1.52	1.53	0.009
SRS Single-Item Predictors					
SRS Single-Item: 9	1,160	.109	.	.	.
SRS Single-Item: 1	1,726	.108	.	.	.
SRS Single-Item: 7	1,129	-.099	.	.	.
SRS Single-Item: 2	1,726	.090	0.55	0.38	-0.187
SRS Single-Item: 4	1,145	-.083	.	.	.
SRS Single-Item: 3	1,160	-.079	3.19	3.35	0.165
SRS Single-Item: 10	1,151	.074	.	.	.
SRS Single-Item: 6					
A vs. B	244	-.072	.	.	.
A vs. C	940	-.051	.	.	.
SRS Single-Item: 12	1,153	-.072	.	.	.
SRS Single-Item: 8					

Table 5.11. (continued)

Predictor	n	r	$M_{Attrits}$	$M_{Nonattrits}$	d
SRS Single-Item Predictors – Continued					
A vs. B	946	.030	.	.	.
A vs. C	521	.056	.	.	.
SRS Single-Item: 16	1,159	-0.050	3.34	3.49	0.105
SRS Single-Item: 11	1,157	0.045	3.93	3.82	-0.095
SRS Single-Item: 15	1,159	-0.030	3.04	3.08	0.063
SRS Single-Item: 13	1,163	0.026	4.16	4.13	-0.054
SRS Single-Item: 14	1,160	-0.026	0.23	0.25	0.054
SRS Single-Item: 5	1,156	-0.020	2.88	2.91	0.042
SRS Single-Item: 17	1,149	-0.018	.	.	.
Non-Operational Demographic Predictors					
Race / Ethnicity					
White vs. Black	1,468	-.138	.	.	.
White vs. Hispanic	1,088	-.105	.	.	.
White vs. Other	1,030	-.057	.	.	.

Note. The coding of all categorical demographic variables presented in this table can be found in the note following Table 5.5. r values reported for categorical variables are phi coefficients. Means, standard deviations, and d statistics were not computed for categorical variables because mean values reflect proportions rather than mean scale scores (see Table 5.10).

Based on Table 5.11, the AIM predictors with the largest difference between those female Tier 1 Soldiers who left the Army and those who did not were AIM Scale C (d = 0.314), the Adaptability Composite (d = 0.300), AIM Scale A (d = 0.260), and AIM Scale F (d = 0.202), with the attrition group mean falling below the nonattrition group mean for all four variables. Minimal differences were apparent between the attrits and nonattrits on both AIM Scales E and G with differences of less than 0.05 standard deviations between the group means.

The operational predictors that exhibited the greatest mean differences between female Tier 1 attrits and nonattrits were the ASVAB Automotive Shop subtest (d = -0.292) and the ASVAB Electronics Information subtest (d = -0.153), with the attrition group mean falling above the nonattrition group mean for both of these variables. Of the other 12 operational predictors, 7 revealed less than 0.1 standard deviation difference between the mean of attrits and the mean of nonattrits.

The SRS predictors with the largest difference between female Tier 1 attrits and nonattrits were Composites 5, 4, and 9, with the attritee mean falling 0.431, 0.416, and 0.302 standard deviations below the nonattritee mean, respectively.

Model Set 1: Evaluating Different AIM Combinations

The first set of models fit to the data for predicting 18-month attrition status among female Tier 1 Soldiers evaluated various combinations of the AIM scales. Table 5.12 presents the result of fitting these models to the female Tier 1 data.

Table 5.12. Comparison of Different AIM Combinations for Predicting Female Tier 1 Soldiers' 18-Month Attrition Status

Model Statistics	AIM Model			
	1	2	3	4
Conditional Odds Ratios				
AIM Adaptability Composite	0.727***	.	0.717***	.
AIM Scale A	.	0.839**	.	0.804**
AIM Scale B	.	1.022	.	1.003
AIM Scale C	.	0.770***	.	0.802***
AIM Scale D	.	.	1.111	1.093
AIM Scale E	.	.	1.186**	1.135*
AIM Scale G	.	.	1.032	1.053
AIM Scale F	.	.	0.846*	0.894
Model-Level Validity				
R	0.145	0.172	0.167	0.182
d	-0.304	-0.362	-0.352	-0.385
Nagelkerke R^2	0.028	0.038	0.037	0.043
AUC	0.576	0.596	0.594	0.605
AUC 95% C.I.	(0.547, 0.604)	(0.568, 0.625)	(0.565, 0.622)	(0.577, 0.633)
Model-Level Utility				
Top 10 % Cut				
PVPT	52.1	55.2	51.2	53.5
False Acceptance Rate	34.8	34.0	34.4	34.1
Top 30 % Cut				
PVPT	45.4	46.0	46.1	45.3
False Acceptance Rate	32.6	31.9	31.8	32.1
Top 50 % Cut				
PVPT	40.8	42.0	42.2	43.8
False Acceptance Rate	32.1	30.1	29.9	28.4

Note. Model 1 contains only the Adaptability Composite (n = 1,705). Model 2 contains each of the Adaptability Composite scales (n = 1,705). Model 3 contains the Adaptability Composite, as well as the four non-Adaptability AIM scales (n = 1,700). Model 4 contains all AIM scales (n = 1,700). * $p < .05$, ** $p < .01$, *** $p < .001$.

Adaptability Composite vs. Adaptability Scales

The first comparison made was between models that examined the predictiveness of the Adaptability Composite by itself versus its component scales. Because these models are not nested, their relative fit to the data could not be compared using differences between their maximized log-likelihood ratios. Therefore, the other criteria described in the methods section were used to compare the relative quality of these models to one another.

Compared to the Tier 2 sample, these two models appear to differ more in both validity and utility for predicting 18-month attrition status among female Tier 1 Soldiers (Model 1: Adaptability Composite, Model 2: Adaptability scales). For example, the point-biserial correlation for the Adaptability Composite model was .145, whereas the correlation for the Adaptability scales model was .172. Although the *AUC* value for the model with the Adaptability scales (0.596) was greater than the *AUC* value for the model with the Adaptability Composite (0.576), this difference did not reach statistical significance.

With regard to the utility of each of these models, depending on the cut score used, the percentage of Soldiers who attrited, given that they were identified as being a likely attrit by the Adaptability Composite model, fell between 40.8% (cutting out the top 50%) and 52.1% (cutting out the top 10%). For the Adaptability scales model these percentages fell between 42% (cutting out the top 50%) and 55.2% (cutting out the top 10%), indicating a slight improvement in PVPT values when the Adaptability scales were entered in separately as predictors. Based on the results presented in Table 5.12, the expected rate of 18-month attrition among female Tier 1 Soldiers if the Adaptability Composite model was adopted would be between 32.1% (cutting out the top 50%) and 34.8% (cutting out the top 10%), compared to expected rates of attrition between 30.1% (cutting out the top 50%) and 34.0% (cutting out the top 10%) if the Adaptability scales model was adopted. Given the base rate of 18-month attrition among female Tier 1 Soldiers in this sample is 36.1%, implementing the cut scores examined here can be expected to reduce attrition among female Tier 1 Soldiers between 3.6% (cutting out the top 10%) and 11.1% (cutting out the top 50%) based on the Adaptability Composite model, and between 5.8% (cutting out the top 10%) and 16.6% (cutting out the top 50%) based on the Adaptability scales model.

Examination of the conditional odds ratios for these first two AIM models reveals the likely cause of the slight improvement in the validity and utility of the Adaptability scales model compared to the Adaptability Composite model. Specifically, AIM Scale B appears to be explaining little variation in attrition relative to Scales A and C. Moreover, those Soldiers who scored higher on Scale B were slightly more likely to attrit (in contrast to the expected directionality: that higher Scale B scores would lead to lower probabilities of attrition), holding the other scales constant. Thus, forming a composite of these three scales based on their sum may lead to a slight decrement in the validity and utility of the prediction model.

Adaptability Composite + Other AIM Scales

Given the differences between the first two models, subsequent comparisons focused on adding to the Adaptability scales model (i.e., Model 2). Nevertheless, a model with the Adaptability Composite and all other AIM scales (i.e., Model 3) was also fitted to the data for comparison.

A model containing the Adaptability scales and the non-Adaptability AIM scales (i.e., Model 4 in Table 5.12) failed to significantly increase model fit over the model containing only the Adaptability scales (χ^2 (4) = 6.80, *n.s.*). Adding the non-Adaptability AIM scales to the Adaptability scales model increased the point-biserial correlation between the AIM model and attrition status only from .172 (Model 2) to .182 (Model 4). Although the *AUC* value for AIM Model 4 (0.605) was higher than the *AUC* value for AIM Model 2 (0.596), the difference was not statistically significant, as the former fell within the latter's 95% confidence interval. Furthermore, the PVPT values and false acceptance rates did not differ appreciably between the two models. Moreover, the PVPT values actually tended to be lower for the present model than those for the Adaptability scales model (at least in the upper ranges of cut scores—e.g., top 30% and up). Lastly, the expected rates of attrition (false acceptance rates) for Model 4 tended to be very similar to those for Model 2. Once again, this was truer for cut scores in the upper ranges of the distribution.

As with the two Adaptability-only models (Models 1 and 2), differences in the validity of Models 3 (Adaptability Composite and other AIM scales) and 4 (all AIM scales) within the female Tier 1 sample were much more visible than in the Tier 2 sample. Although small differences existed between the four AIM models, adding the non-Adaptability scales did appear to improve the fit of the model slightly (this difference was significant only when adding the non-Adaptability AIM scales to the Adaptability Composite (χ^2 (4) = 12.52, $p < .05$).

Several AIM scales were not significant predictors for the attrition status of female Tier 1 Soldiers (i.e., Scale B [Models 2 and 4], Scale G [Models 3 and 4], and Scales F and D [Model 4 only]). Nevertheless, given the significant degree of zero-order relationship observed between these AIM scales and female Tier 1 Soldiers' 18-mon[illegible]trition status (with the exception of Scales E and G), and because the purpose of this rep[illegible] to focus on the AIM, these scales were retained for analysis in subsequent models. Although AIM Models 3 and 4 performed similarly, given the relatively higher performance of the Adaptability Composite when its component scales were entered separately, subsequent model sets examined the impact of adding other predictors to AIM Model 4.

Model Sets 2–4: Adding Extra Predictors to AIM Model 4

The results of adding operational and SRS predictors to the fourth AIM model identified above are presented in Tables 5.13 and 5.14. The format of Table 5.13 is similar to the latter half of Table 5.12 and presents information regarding the validity and utility of each set of models. The predictor variables listed in Table 5.14 are broken down by type (e.g., AIM, operational, SRS) for each set of models, and specific predictor variables within each type are sorted in descending order of their conditional odds ratios.

Note that not all predictor variables initially included in the full models associated with the results presented in Table 5.14 are listed in that table. Only those predictor variables retained in at least one of the models listed in Table 5.14 are presented. Predictor variables that did not meet inclusion criteria for any of the models examined are those that are unlikely to be important predictors of female Tier 1 Soldiers' attrition status.

Table 5.13. Model-Level Comparison of Different Models of Female Tier 1 Soldiers' 18-Month Attrition Status

Model Statistics	AIM Model 4	Set 2: AIM + Operational		Set 3: AIM + Operational + SRS		Set 4: All Predictors	
		Incremental	Best	Incremental	Best	Incremental	Best
Model-Level Validity							
R	0.182	0.266	0.267	0.433	0.434	0.442	0.440
d	-0.385	-0.575	-0.577	-0.999	-1.002	-1.025	-1.020
Nagelkerke R^2	0.043	0.094	0.094	0.249	0.251	0.259	0.256
AUC	0.605	0.675	0.677	0.753	0.754	0.760	0.759
AUC 95% C.I.	(0.577, 0.633)	(0.639, 0.711)	(0.641, 0.712)	(0.721, 0.785)	(0.723, 0.786)	(0.729, 0.792)	(0.727, 0.791)
Model-Level Utility							
Top 10 % Cut							
PVPT	53.5	64.9	65.1	79.6	78.9	77.4	76.3
False Acceptance Rate	34.1	32.9	32.9	31.3	31.4	31.5	31.5
Top 30 % Cut							
PVPT	45.3	52.2	52.0	60.9	61.9	61.3	63.3
False Acceptance Rate	32.1	29.2	29.3	25.5	25.2	25.3	24.4
Top 50 % Cut							
PVPT	43.8	44.8	45.4	52.0	53.1	53.1	53.5
False Acceptance Rate	28.4	27.4	26.8	20.2	19.4	19.1	18.7

Note. Model 4 contains all AIM scales (n = 1,700). Set 2 contains AIM and operational predictors ($n_{Incremental}$ = 1,698, n_{Best} = 1,702). Set 3 contains AIM, operational, and SRS predictors ($n_{Incremental}$ = 930, n_{Best} = 932). Set 4 contains all predictors (non-operational demographics included) ($n_{Incremental}$ = 930, n_{Best} = 934). "Incremental" indicates models fitted using the incremental-fit strategy. "Best" indicates models fitted using the best-composite strategy.

Set 2: AIM + Operational Predictors

When the operational predictors were added to AIM Model 4, the two strategies of model fitting discussed earlier were followed (incremental-fit and best-composite). The reduced Set 2 models were created using both of these strategies, and the results of fitting these models to the female Tier 1 data are discussed in the section that follows.

In the process of reducing the incremental-fit and best-composite models for Set 2, several operational predictors were eliminated (8 for both the incremental-fit and best-composite models). In the process of reducing the best-composite model, two AIM scales (E and G) were eliminated. Both the reduced incremental-fit model ($\Delta\chi2(8) = 4.44$, n.s.) and the reduced best-composite model ($\Delta\chi2(10) = 6.03$, n.s.) provided similar degrees of fit to the data compared to their respective full models. This result emerged despite the fact that the incremental-fit model was based on 8 fewer predictors than its full model, and the best-composite model was based on 10 fewer predictors than its full model (2 AIM predictors, 8 operational predictors). The eliminated variables were correlated with the standardized residuals of the full models from which they were dropped. All resulting correlations were nonsignificant, thus indicating the variables would not likely add incrementally to the validity of either of the reduced models. The variables that were retained in the reduced models of Set 2 and the conditional odds ratios associated with each of them are presented in Table 5.14 under the Set 2 heading.

In fitting these two models to the data, areas of convergence in terms of what predictors appeared to be most important for predicting attrition status were identified. Based on both the incremental-fit and best-composite models, it appears that female Tier 1 Soldiers' who achieved high scores on the ASVAB Automotive Shop subtest ($OR_{\text{Incremental}} = 1.513$; $OR_{\text{Best}} = 1.521$) were more likely to attrit than those who scored low on this test. Conversely, female Tier 1 Soldiers who scored high on AIM Scale A ($OR_{\text{Incremental}} = .771$; $OR_{\text{Best}} = .780$), ASVAB Paragraph Comprehension subtest ($OR_{\text{Incremental}} = .799$; $OR_{\text{Best}} = 0.789$), or AIM Scale C ($OR_{\text{Incremental}} = .832$; $OR_{\text{Best}} = .839$) were less likely to attrit than Soldiers who scored low on these measures.

In terms of the validity and utility of the Set 2 incremental-fit and best-composite models relative to AIM Model 4, adding the operational variables appeared to have a substantial impact. Specifically, when the reduced set of operational variables was added to AIM Model 4 (in the incremental-fit model), there was a significant improvement in model fit ($\Delta\chi^2(6) = 65.82, p < .001$). Similarly, when adding the reduced set of operational variables to the reduced set of AIM Model 4 predictors (in the best-composite model), a significant improvement in the fit of the model also occurred ($\Delta\chi^2(6) = 66.42, p < .001$). Compared to the attrition point-biserial correlation for AIM Model 4 ($r = .182$), both Set 2 models exhibited substantial increases ($r = .266$, incremental-fit model; $r = .267$, best-composite model). The *AUC* values for the two Set 2 models were also significantly greater than the *AUC* value for AIM Model 4. Specifically, the *AUC* value for the Set 2 incremental-fit (0.675) and best-composite models (0.677) both exceeded the upper bound of the 95% confidence interval surrounding the *AUC* value of AIM Model 4 (0.605). The *AUC* values for Set 2 models suggest that if pairs of female Tier 1 Soldiers were randomly selected from those who attrited and those who did not, the Soldier from the attrition group would have a higher score on a composite formed from the Set 2 models approximately two-thirds of the time.

Table 5.14. Predictor Level Comparison of Different Models of Female Tier 1 Soldiers' 18-Month Attrition Status

Predictors	Conditional Odds Ratios					
	Set 2: AIM + Operational		Set 3: AIM + Operational + SRS		Set 4: All Predictors	
	Incremental	Best	Incremental	Best	Incremental	Best
AIM						
AIM Scale D	1.119	1.125	1.201	1.215	1.208	1.183
AIM Scale E	1.145	1.150	1.173	1.176	1.161	
AIM Scale B	1.007		1.062	1.045	1.070	1.085
AIM Scale G	1.065		0.972		0.979	
AIM Scale F	0.846	0.839	1.059	1.048	1.028	
AIM Scale C	0.832	0.839	0.837	0.840	0.825	0.803
AIM Scale A	0.771	0.780	0.736	0.732	0.760	0.803
Operational Predictors						
ASVAB: Automotive Shop	1.513	1.521	1.762	1.759	1.625	1.584
ASVAB: Word Knowledge	1.173	1.168	1.226	1.240	1.202	1.226
ASVAB: Electronics Information	1.156	1.153	1.009		0.977	
ASVAB: Mathematical Knowledge	0.890	0.890	0.976		0.970	
ASVAB: Coding Speed	0.892	0.892	0.887		0.880	0.924
ASVAB: Numerical Operations				0.857		0.893
ASVAB: Paragraph Comprehension	0.799	0.789	0.710	0.716	0.710	0.716
SRS Composite Predictors						
SRS Composite: 13			1.260	1.266	1.289	1.296
SRS Composite: 10			1.214	1.213	1.213	1.230
SRS Composite: 11			1.218	1.227	1.189	1.188
SRS Composite: 21			1.156	1.161	1.144	1.153
SRS Composite: 1			1.123	1.132	1.110	1.114
SRS Composite: 9			0.930	0.919	0.950	0.953
SRS Composite: 18			0.886	0.889	0.874	0.881
SRS Composite: 20			0.877	0.870	0.887	0.895
SRS Composite: 8			0.880	0.879	0.864	0.870
SRS Composite: 4			0.789	0.799	0.797	0.831
SRS Composite: 5			0.779	0.770	0.817	0.828
SRS Composite: 7			0.728	0.723	0.724	0.724
SRS Single-Item Predictors						
SRS Single-Item: 8						
A vs. B			1.230	1.228	1.188	1.201
A vs. C			1.311	1.273	1.300	1.370
SRS Single-Item: 9			1.259	1.271	1.292	1.321
SRS Single-Item: 1			1.286		1.341	1.183
SRS Single-Item: 15			1.210	1.197	1.196	
SRS Single-Item: 2				1.085		1.068
SRS Single-Item: 14			0.833	0.826	0.813	0.815
SRS Single-Item: 7			0.806	0.796	0.783	0.765
SRS Single-Item: 12			0.735	0.726	0.717	0.713

Table 5.14. (continued)

Predictors	Conditional Odds Ratios					
	Set 2		Set 3		Set 4	
	Incremental	Best	Incremental	Best	Incremental	Best
SRS Single-Item: 4			0.633	0.617	0.635	0.640
SRS Single-Item: 6						
A vs. C			0.436	0.431	0.430	0.467
A vs. B			0.415	0.403	0.404	0.426
Non-Operational Demographic Predictors						
Race / Ethnicity						
White vs. Black					0.592	0.566
White vs. Hispanic					0.537	0.499
White vs. Other					0.845	0.804

Note. The coding of all categorical demographic variables presented in this table can be found in Table 5.5. Set 2 contains AIM and operational predictors. Set 3 contains AIM, operational, and SRS predictors. Set 4 contains all predictors (non-operational demographics included). "Incremental" indicates models fit using the incremental-fit strategy. "Best" indicates models fit using the best-composite strategy.

In terms of the quality of decisions made based on the Set 2 models compared to AIM Model 4, improvement was once again apparent. Focusing only on the PVPT values when cutting out the top 10% of female Tier 1 Soldiers, the Set 2 models resulted in 11.6% (best-composite model) and 11.4% (incremental-fit model) percentage point increases compared to AIM Model 4. Also apparent was the impact that adding the operational variables had on the expected female Tier 1 attrition rate. Namely, including the operational predictors identified by the Set 2 models and cutting out the top 10% of female Tier 1 Soldiers based on the resulting composite would be expected to yield an 8.9% decrease in attrition relative to the female Tier 1 base rate of 36.1%. Note that this is compared to a 5.5% improvement on the base rate of attrition among female Tier 1 Soldiers with AIM Model 4 using a top 10% cut.

Set 3: AIM + Operational Predictors + SRS Predictors

As was the case with the Set 3 models examined in the Tier 2 sample, the incremental-fit and best-composite modeling strategies were employed in fitting the current set of models for Tier 1 females. In the process of reducing the incremental-fit and best-composite models, several SRS predictors were eliminated (17 for the incremental-fit model, 17 for the best-composite model). Both the reduced incremental-fit Set 3 model ($\Delta\chi^2(17) = 5.67$, *n.s.*) and the reduced best-composite Set 3 model ($\Delta\chi^2(28) = 5.55$, *n.s.*) provided similar degrees of fit to the data compared to their respective full models. This was so even though the incremental-fit model contained 17 fewer predictors than its full model, and the best-composite model was based on 28 fewer predictors than its full model (1 AIM predictor, 10 operational predictors, and 17 SRS predictors were eliminated). The eliminated variables were correlated with the standardized residuals of the full models from which they were dropped. All resulting correlations were nonsignificant, thus indicating the variables would not likely add incrementally to the validity of either of the reduced models. The variables that were retained in the reduced models of Set 3 and the conditional odds ratios associated with each of them are presented in Table 5.14 under the Set 3 heading.

In fitting these two models to the data, we identified areas of convergence in terms of what factors appeared to be most important for predicting attrition status. As was the case with the Set 2 models, based on both the incremental-fit and best-composite models, female Tier 1 Soldiers' who scored high on the ASVAB Automotive Shop subtest ($OR_{Incremental}$ = 1.762; OR_{Best} = 1.759) were more likely to attrit than those who scored low on that subtest. Conversely, female Tier 1 Soldiers who scored high on the ASVAB Paragraph Comprehension subtest ($OR_{Incremental}$ = .710; OR_{Best} = 0.716) were less likely to attrit than those who scored low on that subtest. With regard to the SRS predictors, as was the case with Tier 2 Soldiers, female Tier 1 Soldiers who scored in category B ($OR_{Incremental}$ = 1.230; OR_{Best} = 1.228), or C ($OR_{Incremental}$ = 1.311; OR_{Best} = 1.273) on SRS Single-Item Predictor 8 were more likely to attrit than those who scored in category A. Conversely, Tier 1 females who scored in category B ($OR_{Incremental}$ = .415; OR_{Best} = .403) or C ($OR_{Incremental}$ = .436; OR_{Best} = .431) on SRS Single-Item Predictor 6 were less likely to attrit than those scored in category A. The effects of item 6 appeared to be much stronger among female Tier 1 Soldiers than it was among Tier 2 Soldiers. Additionally, female Tier 1 Soldiers who endorsed SRS Single-Item Predictor 4 were more likely to attrit ($OR_{Incremental}$ = .633; OR_{Best} = .617) than those who did not endorse the item.

One of the primary reasons for fitting the best-composite model in Set 3 was to see whether any of the AIM or operational variables retained in the Set 2 reduced models would be eliminated when SRS variables were added to the model. No new AIM scales were eliminated as a result of adding the SRS predictors. In fact, the only AIM scale that failed to meet inclusion criteria for the Set 3 best-composite model was Scale G. With regard to the ASVAB subtests retained in the Set 2 models, however, three were eliminated from the best-composite model in Set 3 (Electronics Information, Mathematical Knowledge, and Coding Speed). These results suggest that the variance in attrition being tapped by these variables is being accounted for by the SRS variables.

In terms of the validity and utility of the Set 3 incremental-fit and best-composite models relative to models containing just the AIM and operational variables, adding the SRS variables had a substantial impact. Specifically, when the reduced set of SRS variables from the Set 3 incremental-fit model was added to the Set 2 incremental-fit model, there was a significant improvement in the fit of the model to the data ($\Delta\chi^2(23) = 86.10, p < .001$). Similarly, when adding the reduced set of SRS variables to the reduced set of AIM and operational variables, in the third step of the Set 3 best-composite model, a significant improvement in fit of the model occurred ($\Delta\chi^2(23) = 87.58, p < .001$). Compared to the attrition point-biserial correlation for the Set 2 models (r = .266, incremental-fit model; r = .267, best-composite model), both Set 3 models exhibited substantial increases in validity (r = .433, incremental-fit model; r = .434, best-composite model). The *AUC* values for the two Set 3 models were also significantly greater than the *AUC* value for the Set 2 models. Specifically, the *AUC* value for the Set 3 incremental-fit model (0.753), as well as the *AUC* value for the Set 3 best-composite model (0.754) both exceeded the upper bound of the 95% confidence interval surrounding the Set 2 model's *AUC* values ($AUC_{Incremental}$ = 0.675; AUC_{Best} = 0.677). The *AUC* values for the Set 3 models suggest that if pairs of female Tier 1 Soldiers were randomly selected from those who attrited and those who did not, the Soldier from the attrition group would have a higher score on a composite formed from the Set 3 models approximately 75% of the time.

In terms of the quality of decisions made based on the Set 3 models compared to the Set 2 models, improvement was again apparent. Focusing only on the PVPT values when cutting out the top 10% of female Tier 1 Soldiers, the Set 3 models resulted in 13.8 (best-composite model) and 14.7 (incremental-fit model) percentage point increases compared to their respective Set 2 models. Adding the SRS variables also had an apparent impact on the expected female Tier 1 attrition rate. Namely, including the SRS predictors identified by the Set 3 models and cutting out the top 10% of female Tier 1 Soldiers based on the resulting composite would be expected to yield a 13.3% (incremental-fit model false acceptance rate = 31.3%) or 13.0% (best composite-model false acceptance rate = 31.4%) decrease in attrition relative to the female Tier 1 base rate of 36.1%. Note that this is compared to an approximately 8.9% improvement on the base rate of attrition among female Tier 1 Soldiers when using a top 10% cut with the AIM and operational variables only (i.e., the Set 2 models).

Set 4: Demographics + AIM + Operational Predictors + SRS Predictors

The final set of female Tier 1 models examined the impact of entering non-operational demographics (race was the only demographic variable entered for female Tier 1 Soldiers) into the attrition model on the results obtained from the predictors retained by the models in Set 3. In reducing the Set 4 best-composite model, 29 variables were eliminated (3 AIM predictors, 9 operational predictors, and 17 SRS predictors). Even after eliminating the 29 predictors, this model provided similar fit to the data relative to the model with all predictors entered ($\Delta\chi^2(29) = 9.32$, *n.s.*). The eliminated variables were correlated with the standardized residuals of the reduced model from which they had been dropped. Again, all correlations were nonsignificant, indicating that the eliminated variables would be unlikely to add incrementally to the validity of the reduced model. The variables retained for the Set 4 models and the conditional odds ratios associated with each of them are presented in Table 5.14 under the Set 4 heading.

In fitting both the Set 4 models to the data, areas of convergence in terms of what predictors appeared to be most important for predicting attrition status were identified. As was the case with the Set 3 models, based on both the incremental-fit, and best-composite models the two strongest predictors of attrition once again dealt with Soldiers' scores on the Automotive Shop subtest of the ASVAB and their plans for college after their enlistment term was up. Specifically, female Tier 1 individuals who scored high on the ASVAB Automotive Shop subtest ($OR_{\text{Incremental}} = 1.625$; $OR_{\text{Best}} = 1.584$) were more likely to attrit than those who scored low. Conversely, female Tier 1 Soldiers who scored in category B ($OR_{Incremental} = .404$; $OR_{\text{Best}} = .426$) or C ($OR_{\text{Incremental}} = .430$; $OR_{\text{Best}} = .467$) on SRS Single-Item Predictor 6 were less likely to attrit than those who scored in category A.

One of the primary reasons for examining the Set 4 models was to see if any predictors retained in the Set 3 best-composite model would be eliminated if demographics were first considered in the model. AIM Scales E and F both failed to reach criteria for inclusion when race was given consideration in the Set 4 best-composite model. For the most part, however, adding race had very little impact on the other predictors in the model, as most variables that were retained in the Set 3 best-composite were also retained in the Set 4 best-composite model with very little change in their respective conditional odds ratio. Given that race cannot (and would not) be used to select Soldiers, the impact that race had on the overall validity and utility relative to the other sets of models will not be detailed here. Nevertheless, it is worth noting that the validity and utility of the Set 4 models appears to be very similar to their Set 3 counterparts, suggesting that including the race variable would not likely make a large difference on the validity or utility of an overall composite of attrition predictors.

Female Tier 1 Attrition Model Summary

Figure 5.2 provides a graphical summary of the utility of the female Tier 1 attrition models examined in the present investigation. Specifically, Figure 5.2 displays the percentage of female Tier 1 recruits that attrited within each decile of the predictor composite stemming from the reduced models for each set of analyses performed. Given the similarity of the best-composite and incremental-fit models within each set of analyses, only the best-composite models were plotted in Figure 5.2.

As with the Tier 2 sample, the results for Tier 1 females suggest that the Army may have much to gain by further exploring the SRS and operational variables identified here as potentially salient predictors of female Tier 1 Soldiers' 18-month attrition status (potential for about a 13% decrease in female Tier 1 attrition with a top 10% cut). Moreover, unlike the results for the Tier 2 sample, substantial gains also appear possible when only operational variables are added to the AIM (potential for about an 9% decrease in attrition with a top 10% cut).

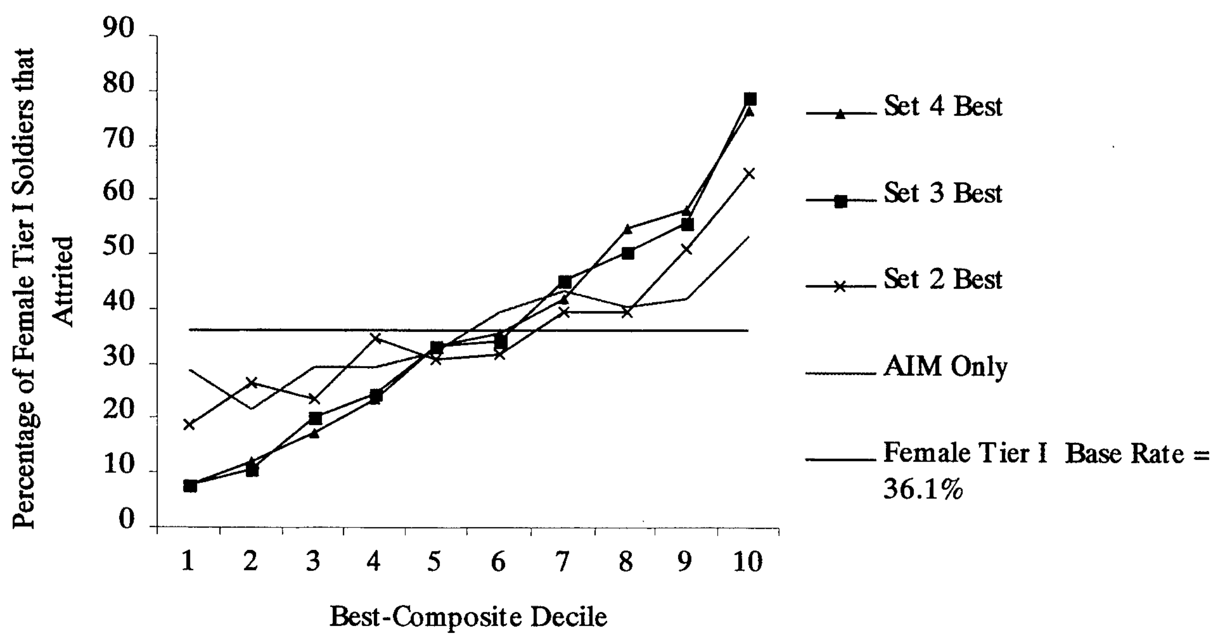


Figure 5.2. Observed percentage of female Tier 1 Soldiers that attrited at or before 18 months by best-composite model decile.

One of the more interesting findings that the present investigation of female Tier 1 Soldiers revealed was that the AIM Scale B component of the Adaptability Composite appeared to be positively related to attrition when other Adaptability components were held constant. This is in contrast to the other Adaptability Composite components, which both appeared to be negatively related to attrition (and held a greater magnitude as well) when holding other predictors constant. These findings suggest that forming an Adaptability composite based on the sum of the current three components will weaken the criterion-related validity of the composite. The current scoring of the Adaptability Composite exacerbates the problem because AIM Scale B receives more weight than either Scale C or A. These latter scales exhibited much stronger relationships (and in the expected direction) to the attrition criteria than Scale B.

As was the case with Tier 2 Soldiers, female Tier 1 Soldiers' standing on SRS Single-Item Predictor 5 appeared to be one of the most salient predictors of attrition (in the case of female Tier 1 Soldiers it was the strongest predictor). Among the operational predictors, the ASVAB Automotive shop subtest was most positively related to female Tier 1 Soldier attrition. One potential explanation for the significance of this effect is that high scores on this test may place female recruits into more mechanically oriented MOS. Such mechanically oriented MOS have traditionally been populated by males. Thus, Tier 1 female Soldiers given such assignments may feel undue pressures due to underrepresentation of females in such MOS. Alternatively, a simpler explanation for this finding might be that females scoring high on this ASVAB subtest might lack interest in the MOS to which they were assigned. Thus, the ASVAB Automotive Shop subtest may actually be serving as a proxy for MOS, which may be predictive of female Tier 1 attrition status.

In sum, future research with regard to female Tier 1 Soldiers' attrition status should:

- Investigate the efficacy of adopting different scoring routines for the Adaptability Composite that give less weight to, or even exclude, AIM Scale B.

- Determine if the variables identified in the present investigation as potentially salient predictors of 18-month attrition status cross-validate well in other samples of female Tier 1 Soldiers.

- Establish a more solid rationale for why the variables identified are predictive of attrition (e.g., investigate between-MOS attrition rate differences as a potential explanation for the significance of the ASVAB Automotive Shop subtest).

- Identify which of the salient SRS predictors might be used operationally during the enlistment process to help identify female Tier 1 recruits who are most at risk of attrition.

- Subsequently, test the potentially operational SRS predictors to see if their addition to AIM and operational variables results in reductions in attrition similar to those found in the present investigation.

Discussion

One of the primary purposes of the present investigation was to identify predictors that could be used operationally to supplement the AIM and increase the validity and utility of predicting Soldiers' 18-month attrition status. To achieve this end, sets of currently available operational variables were examined, as well as variables obtained from the Soldier Reception Survey. Working from a full set of 14 operational predictors and 38 SRS predictors, models were fitted to the attrition data of both Tier 2 and female Tier 1 Soldiers. Upon reducing these full models to those variables that met liberal inclusion criteria, 5 of the 14 operational variables remained for the Tier 2 Soldiers (Set 3 best-composite) and only 4 of the 14 remained for the female Tier 1 Soldiers. Only one operational variable appeared in the reduced models for each sample (ASVAB Word Knowledge). With regard to the 38 SRS variables, 17 of the 38 variables remained for the Tier 2 Soldiers (Set 3 best-composite) and 21 of the 38 variables remained for the female Tier 1 Soldiers. Unlike the operational variables, 14 of the SRS variables were identified as salient for predicting both Tier 2 and female Tier 1 Soldiers' attrition status.

Considering the number of SRS variables identified as salient predictors of 18-month attrition, it is likely that some degree of redundancy among the variables exists. Future research can likely eliminate such redundancy by addressing a very practical question: Of the SRS variables identified in the present investigation as potentially salient predictors of Soldiers' attrition status, which may be successfully used operationally? One problem that arises with using some of the SRS variables found to predict attrition is that their validity may not hold up if they are implemented as part of an operational screen. Some SRS variables are very transparent and others evoke unverifiable responses (e.g., "How confident are you that you will be successful in the Army?"). Given these characteristics of many of the SRS items, they may be highly susceptible to faking if administered as part of an operational screen. One potential strategy to address these concerns may be to try to develop operationally viable measures of the constructs that the highly salient SRS predictors identified in the present investigation. Ideally such measures would be less transparent and more verifiable than their SRS counterparts. By retaining only operationally viable SRS predictors, the resulting set of SRS predictors may be much smaller than the 17 (Tier 2) to 21 (Tier 1 females) SRS variables identified in the present investigation.

Future investigations also might focus on potential non-linear relationships between each predictor examined in the present investigation and Soldiers' 18-month attrition status. The present investigation focused only on examining linear relationships among the set of predictor variables and attrition. Linear relationships mean that a given predictor is equally predictive of attrition across its entire range of scores. Nonlinear relationships, on the other hand, mean that a given predictor becomes more or less predictive of attrition across its range of scores. Although such a check for potential nonlinearities is worthy of future attention, such checks were not given high priority in the present investigation because the linear approximation of relationships between two or more variables is typically robust for even moderate nonlinearities in those relationships.

With regard to the AIM, Scales C and A consistently emerged as the most predictive of attrition. Nevertheless, the manner in which the AIM Adaptability Composite is currently formed gives more weight to Scale C (contributing up to 40 points) than to either Scale C (contributing up to 16 points) or Scale A (contributing up to 38 points). Based on the results of the present investigation, it appears there is a need to reevaluate how the components of the Adaptability Composite are combined (e.g., allowing the scales to have equal weight, or giving relatively more weight to Scales C or A).

CHAPTER 6. AIM ADVERSE IMPACT, DIFFERENTIAL VALIDITY, AND DIFFERENTIAL PREDICTION

Paul R. Sackett and Roxanne M. Laczo
University of Minnesota

The work reported in this chapter was carried out to ensure that AIM, as it is currently being used as an attrition screen under the GED Plus program, adequately meets both legal and professional guidelines with respect to adverse impact. In addition, the authors provide an in depth analysis of differential validity and differential prediction by gender and race. In more recent efforts, outside the timeframe of this report, ARI examined the potential adverse impact of new attrition screening measures that were proposed as replacements for the existing AIM Adaptability Composite. The level of effort required for these new evaluations was minimal, because they were strictly limited to the question of adverse impact.

This chapter addresses issues in the broad domain of "fairness." We investigated three issues: (a) subgroup differences and adverse impact, (b) differential validity of AIM by gender and race in predicting attrition, and (c) differential prediction (or predictive bias) by gender and race in predicting attrition.

Subgroup Differences and Adverse Impact

We examined subgroup mean differences on the AIM Adaptability Composite by race and gender. Adverse impact in a selection procedure is a function of two things: (a) a measurement property of the selection procedure (i.e., the magnitude of the mean difference between groups) and (b) an administrative decision as to where in the selection procedure the cutoff is set.

Table 6.1 presents subgroup means, standard deviations, and standardized mean differences (*d*) by gender and by race for the research datasets of the Army and the Air Force, and for the Army operational dataset. All mean differences between groups are small. Women score slightly higher than men in the Army research sample and slightly lower than men in the Army operational sample and the Air Force sample. Blacks and Hispanics score equal to or slightly higher than Whites in all samples. American Indians score slightly lower than Whites in the Army research sample and slightly lower than Whites in the Army operational sample, while Asians score slightly higher than Whites in the Army research sample and lower in the operational Army sample; data from these groups are not reported in the Air Force sample.

The question of interest is whether subgroup differences of this magnitude could translate into adverse impact against any subgroup. The most common approach to adverse impact is the four-fifths, or 80%, rule of thumb put forward in the *1978 Uniform Guidelines on Employee Selection Procedures*, issued by the EEOC and other federal agencies. This rule suggests adverse impact when the selection ratio for the lowest performing subgroup is less than 80% of the selection ratio for the highest performing subgroup.

Sackett and Ellingson (1997) offer a table that shows the value of the four-fifths ratio for various combinations of subgroup differences (*d*) and majority group selection ratios (with "majority group" in actuality defined as "highest scoring subgroup"), assuming normality and homogeneity of subgroup variances. The largest *d* in the Army/Air Force data reported here in Table 6.1 is the 0.17 value for gender in Air Force data. The closest tabled value in Sackett and Ellingson is 0.20. Sackett and Ellingson show that a *d* as small as 0.20 can result in adverse impact, but only at selection ratios of less than 50%. As the AIM is being evaluated for use in a "screen-out" fashion, with relatively high selection ratios, *d*s in the 0.1–0.2 range will not produce adverse impact. At the projected AIM selection ratio of 75% (e.g., a score above the 25th percentile required for selection), a *d* of 0.2 produces an adverse impact ratio of 0.91; a *d* of 0.1 produces an adverse impact ratio of 0.95.

Table 6.1. Subgroup Differences: Adaptability Composite

	Army Research				Army Operational				Air Force			
	n	*M*	*SD*	*d*	*n*	*M*	*SD*	*d*	*n*	*M*	*SD*	*d*
Gender												
Male	19,143	56.03	10.39	--	6,145	63.78	8.18	--	14,172	58.75	10.37	--
Female	3,357	56.56	10.10	-0.05	826	63.37	8.07	0.05	5,129	56.99	10.18	0.17
Total	22,493	56.11	10.34	--	6,971	63.73	8.17	--	19,301	58.28	10.35	--
Race												
White	13,522	55.61	10.87	--	5,296	63.79	8.22	--	11,447	57.97	10.71	--
Black	4,765	57.18	9.30	-0.15	719	63.69	8.12	0.01	3,087	58.83	9.65	-0.08
Hispanic	2,683	57.02	9.27	-0.13	629	63.79	7.51	0.00	1,141	59.32	9.36	-0.13
Amer. Ind.	322	54.73	10.77	0.08	102	62.87	8.37	0.11	--	--	--	--
Asian	745	55.99	9.69	-0.04	111	62.41	7.72	0.17	--	--	--	--
Other	456	55.64	10.40	0.00	102	62.98	9.55	0.10	1,013	58.40	9.95	-0.04
Total	22,493	56.11	10.34	--	6,959	63.74	8.17	--	16,688	58.25	10.40	--

Note. Effect sizes represent male mean minus female mean; White mean minus other mean. Eskimo excluded from all analysis because n=7. American Indian, Eskimo, and Asian data for Air Force contained in Other category.

Thus, the AIM produces very small subgroup differences, none of which result in adverse impact using the four-fifths rule when the AIM is used with a low cutoff in order to screen out low-scoring individuals. Note that these findings result from administering the AIM in both operational and nonoperational settings. While scores are much higher in operational settings, subgroup differences remain essentially the same.

Differential Validity

This section of the report presents the relationship between AIM Adaptability Composite scores and attrition at various time periods. The time periods involved range from 3 to 27 months for both the Army and the Air Force Research data, in 3-month intervals. Only the two research datasets were included in these analyses, since only limited attrition data were available for the operational sample. The standardized mean difference (*d*) between the AIM scores of individuals who remain in the service (nonattritees) and those who leave the service (attritees) is used to index this relationship. This *d* index is computed as attritee mean minus nonattritee mean, divided by the pooled within-group standard deviation. Thus a negative value indicates higher AIM scores for nonattritees. Findings are presented for both the Army and the Air Force research datasets, and are reported separately for each racial and gender group.

Validity and differential validity statistics are reported in two different ways here, reflecting different choices about how to examine the data. Both merit examination to obtain a full picture of the relationship between AIM and attrition. Each set of analyses results in a *d* value for each subgroup for each time period. The *d* statistic is used to index the size of the mean AIM difference between nonattritees and attritees; it is used instead of a correlation metric (e.g., point-biserial correlations) because *d* is not influenced by differences in subgroup proportions (i.e., there are generally far more nonattritees than attritees).

Individuals in the Army AIM research dataset entered the service at different points in time. The earliest entrants had the potential to have been in the service for 27 months at the point when the data were finalized for analysis; the latest entrants had the potential to have been in the service for 18 months at the point where the data were finalized for analysis. The first set of analyses includes only individuals with the potential to have been in the service for the amount of time in question. Thus the analysis for the 21-month time period includes only individuals who could potentially have 21 months of service. For example, an individual who entered the service 18 months before the dataset was finalized would not be included in the analyses, even if that individual left the service after 6 months. The rationale here is that the 21-month attrition status of individuals with less than 21 months of service potential cannot be known for all individuals. Thus these analyses focus on the cohort of individuals with the potential to have been in the service for the amount of time specified.

Tables 6.2 and 6.3 focus on the first approach outlined above, with Table 6.2 focusing on the Army research dataset and Table 6.3 on the Air Force research dataset. A number of conclusions can be drawn from the tables. First, focusing on the initial 3-month period, the overall *d* is slightly higher for the Air Force (-0.53) than for the Army (-0.40). Second, while the Army data show virtually identical effect sizes for males and females, the Air Force data show higher validity for males (-0.55) than females (-0.46).

Third, both the Army and the Air Force data show some variability in the validity values for various racial subgroups. We caution against overinterpreting these differences: The stability of the *d* values is heavily influenced by the sample size in the attrition group, and the number of attrits for some groups is very small (e.g., 36 for American Indians and 33 for Asian-Americans in the Army data). Table 6.4 shows the results of tests of the significance of the difference between *d* values for males vs. females, and for each racial subgroup in comparison to the White group. None of the male-female differences were significant in either the Army or the Air Force data. In the Army data, the racial subgroup comparisons generally showed no significant differences through 18 months, which is the last time period for which complete data are available. The Black-White difference was significant at 9 months, but not at other time periods. In contrast, in the Air Force data the Black-White difference was consistently significant across time periods. In these cases, validity was higher for white Soldiers.

Table 6.2. Differential Validity Effect Sizes: Cumulative Data for Army

		Attrition Period								
		3-Month	6-Month	9-Month	12-Month	15-Month	18-Month	21-Month	24-Month	27-Month
Gender										
Male	*d*	-0.41	-0.37	-0.36	-0.35	-0.34	-0.33	-0.34	-0.36	-0.39
	Nonattrit *n*	17,422	16,372	15,968	15,679	15,358	15,000	11,861	5,988	3,725
	Attrit *n*	1,717	2,745	3,127	3,408	3,726	4,083	3,725	1,950	1,218
Female	*d*	-0.42	-0.41	-0.39	-0.37	-0.34	-0.35	-0.35	-0.28	-0.34
	Nonattrit *n*	2,713	2,482	2,386	2,294	2,208	2,094	1,692	914	686
	Attrit *n*	644	874	965	1,054	1,139	1,250	1,132	705	543
Race										
White	*d*	-0.40	-0.38	-0.36	-0.35	-0.34	-0.33	-0.35	-0.35	-0.37
	Nonattrit *n*	11,879	11,009	10,671	10,424	10,159	9,859	7,717	3,788	2,398
	Attrit *n*	1,640	2,489	2,807	3,044	3,307	3,604	3,272	1,700	1,102
Black	*d*	-0.36	-0.30	-0.27	-0.27	-0.26	-0.27	-0.28	-0.27	-0.29
	Nonattrit *n*	4,345	4,125	4,045	3,973	3,896	3,807	3,083	1,665	1,077
	Attrit *n*	419	639	717	788	863	951	860	524	369
Hispanic	*d*	-0.31	-0.28	-0.28	-0.27	-0.24	-0.23	-0.20	-0.19	-0.28
	Nonattrit *n*	2,508	2,390	2,344	2,299	2,257	2,207	1,779	951	627
	Attrit *n*	175	291	334	379	421	471	432	249	164

Table 6.2. (continued)

		Attrition Period								
		3-Month	6-Month	9-Month	12-Month	15-Month	18-Month	21-Month	24-Month	27-Month
Amer. Ind.	*d*	-0.27	-0.24	-0.30	-0.30	-0.20	-0.22	-0.23	-0.28	-0.54
	Nonattrit *n*	286	267	258	255	249	240	185	88	49
	Attrit *n*	36	55	64	67	73	82	72	43	26
Asian	*d*	-0.35	-0.41	-0.39	-0.32	-0.27	-0.28	-0.10	-0.24	-0.23
	Nonattrit *n*	712	690	676	667	657	648	523	262	166
	Attrit *n*	33	55	69	78	88	97	96	66	42
Other	*d*	-0.41	-0.42	-0.43	-0.44	-0.45	-0.46	-0.41	-0.27	-0.47
	Nonattrit *n*	398	366	353	348	341	326	260	144	93
	Attrit *n*	58	90	101	106	113	128	125	73	58
Total	*d*	-0.40	-0.37	-0.36	-0.34	-0.33	-0.32	-0.33	-0.32	-0.36
	Nonattrit *n*	20,135	18,854	18,354	17,973	17,566	17,094	13,553	6,902	4,411
	Attrit *n*	2,361	3,619	4,092	4,462	4,865	5,333	4,857	2,655	1,761

Note. Effect sizes represent Attrit mean minus Nonattrit mean.

Table 6.3. Differential Validity Effect Sizes: Cumulative Data for Air Force

		Attrition Period								
		3-Month	6-Month	9-Month	12-Month	15-Month	18-Month	21-Month	24-Month	27-Month
Gender										
Male	*d*	-0.55	-0.48	-0.45	-0.44	-0.42	-0.41	-0.39	-0.38	-0.36
	Nonattrit *n*	13,148	12,889	12,748	12,674	12,610	12,540	12,453	12,354	12,286
	Attrit *n*	1,024	1,283	1,424	1,498	1,562	1,632	1,719	1,818	1,785
Female	*d*	-0.46	-0.40	-0.37	-0.36	-0.36	-0.35	-0.34	-0.32	-0.29
	Nonattrit *n*	4,639	4,559	4,511	4,492	4,463	4,430	4,391	4,359	4,327
	Attrit *n*	490	570	618	637	666	699	738	770	756
Race										
White	*d*	-0.55	-0.50	-0.46	-0.45	-0.44	-0.44	-0.42	-0.41	-0.39
	Nonattrit *n*	10,300	10,082	9,959	9,901	9,845	9,783	9,707	9,621	9,562
	Attrit *n*	1,147	1,365	1,488	1,546	1,602	1,664	1,740	1,826	1,776
Black	*d*	-0.33	-0.29	-0.28	-0.27	-0.27	-0.24	-0.22	-0.20	-0.18
	Nonattrit *n*	2,838	2,752	2,703	2,679	2,655	2,628	2,593	2,567	2,541
	Attrit *n*	249	335	384	408	432	459	494	520	520
Hispanic	*d*	-0.63	-0.46	-0.44	-0.42	-0.40	-0.38	-0.37	-0.36	-0.30
	Nonattrit *n*	1,080	1,056	1,047	1,042	1,036	1,030	1,021	1,015	1,007
	Attrit *n*	61	85	94	99	105	111	120	126	129

Table 6.3. (continued)

		Attrition Period								
		3-Month	6-Month	9-Month	12-Month	15-Month	18-Month	21-Month	24-Month	27-Month
Other	*d*	-0.61	-0.50	-0.44	-0.49	-0.51	-0.56	-0.56	-0.56	-0.52
	Nonattrit *n*	962	953	947	942	937	929	923	918	913
	Attrit *n*	51	60	66	71	76	84	90	95	96
Total	*d*	-0.53	-0.47	-0.43	-0.42	-0.41	-0.40	-0.39	-0.37	-0.35
	Nonattrit *n*	17,787	17,448	17,259	17,166	17,073	16,970	16,844	16,713	16,613
	Attrit *n*	1,514	1,853	2,042	2,135	2,228	2,331	2,457	2,588	2,541

Note. Effect sizes represent Attrit mean minus Nonattrit mean.

Table 6.4. Z-test for Difference Between Differential Validity Effect Sizes

		Attrition Period								
		3-Month	6-Month	9-Month	12-Month	15-Month	18-Month	21-Month	24-Month	27-Month
Army	Male-Female	0.20	0.89	0.69	0.48	0.00	0.50	0.23	-1.41	-0.75
	White-Black	-0.69	-1.66	-1.96*	-1.81	-1.87	-1.45	-1.59	-1.37	-1.13
	White-Hispanic	-1.09	-1.51	-1.28	-1.35	-1.76	-1.84	-2.60*	-2.07*	-0.94
	White-Amer. In.	-0.73	-0.93	-0.42	-0.36	-1.04	-0.85	-0.85	-0.37	0.68
	White-Asian	-0.28	0.21	0.23	-0.25	-0.61	-0.45	-2.21*	-0.78	-0.79
	White-Other	0.07	0.33	0.60	0.79	0.99	1.21	0.54	-0.54	0.58
Air Force	Male-Female	-1.56	-1.50	-1.56	-1.58	-1.21	-1.23	-1.05	-1.29	-1.49
	White-Black	-3.00*	-3.24*	-2.93*	-3.00*	-2.90*	-3.49*	-3.59*	-3.85*	-3.84*
	White-Hispanic	0.59	-0.34	-0.18	-0.28	-0.38	-0.58	-0.50	-0.51	-0.93
	White-Other	0.41	0.00	-0.15	0.32	0.57	1.02	1.23	1.35	1.17

Note. Based on values taken from Table 6.2 for Army and Table 6.3 for Air Force.
*$p < .05$.

Fourth, there is a general trend toward decreasing validities across time periods. Insight into this finding will be provided when the second set of differential validity analyses (i.e., noncumulative vs. cumulative analyses) are presented below.

Tables 6.5 and 6.6 present the second set of differential validity analyses for the Army and Air Force research datasets respectively. These analyses take a noncumulative perspective on attrition. The *d* values for each time period are based on a comparison of the mean Adaptability Composite score of individuals still in the service with the mean of those individuals leaving in the 3-month period since attrition was last assessed. In other words the *d* value for 6 months is based on comparing the AIM mean of individuals in the service after 6 months with the mean score of individuals leaving the service between month 3 and month 6. Thus this analysis permits the determination of whether AIM continues to differentiate nonattritees and attritees at various points through the first term of enlistment.

The clear message emerging from these tables is that validity is highest for the initial 3 months of service, and declines quite dramatically after that. Because the number of individuals leaving the service in each time period was quite small for some subgroups, data were collapsed into three time periods: the first 3 months, months 4–12 (the remainder of the first year of service), and months 13–24 (the second year of service). The total sample analysis reveals that in the Army sample, validity drops from -0.40 to -0.20 to -0.11 across the three time periods; in the Air Force sample, validity drops from -0.53 to -0.11 to -0.06 across the same three time periods.

Table 6.7 presents results by subgroup. Results by gender are presented graphically in Figures 8.1 and 8.2; results by race are presented in Figures 8.3 and 8.4. Males, females, Whites, and Hispanics show the same pattern of results; the Black subgroup, in contrast, shows AIM validity increasing in year 2, a pattern not replicated in the Air Force data.

Differential Prediction

Analysis of Army and Air Force Data

Investigating differential prediction by race and gender in the use of tests or other assessment instruments for personnel selection has been a long-standing concern for both researchers and practitioners. Differential prediction is commonly assessed using the regression model proposed by Cleary (1968). This model tests the within-group regression lines relating test scores to a job-relevant criterion for differences in slopes, intercepts, and sometimes error variances. No predictive bias exists if the predictive relationship in the two groups being compared can be described by a common regression line. Differences in slopes or intercepts would imply bias, because systematic errors of prediction would be made on the basis of group membership. Both the *AERA/APA/ NCME Standards for Educational and Psychological Testing* (1999) and the SIOP *Principles for the Validation and Use of Personnel Selection Procedures* (1987) acknowledge this as the accepted approach to examining predictive bias.

Table 6.5. Differential Validity Effect Sizes: Noncumulative Data for Army

		Attrition Period								
		3-Month	6-Month	9-Month	12-Month	15-Month	18-Month	21-Month	24-Month	27-Month
Gender										
Male	*d*	-0.41	-0.29	-0.25	-0.18	-0.18	-0.20	-0.31	-0.19	-0.30
	Nonattrit *n*	17,422	16,372	15,968	15,679	15,358	15,000	11,861	5,988	3,725
	Attrit *n*	1,716	1,007	391	277	320	342	374	226	117
Female	*d*	-0.42	-0.31	-0.17	-0.14	0.05	-0.34	-0.18	-0.05	-0.07
	Nonattrit *n*	2,713	2,482	2,386	2,294	2,208	2,094	1,692	914	686
	it *n*	644	227	91	90	83	108	81	60	34
Race										
White	*d*	-0.40	-0.30	-0.24	-0.13	-0.17	-0.20	-0.27	-0.15	-0.04
	Nonattrit *n*	11,879	11,009	10,671	10,424	10,159	9,859	7,717	3,788	2,398
	Attrit *n*	1,639	834	324	232	267	284	303	175	88
Black	*d*	-0.36	-0.19	-0.02	-0.23	-0.06	-0.29	-0.49	-0.23	-0.43
	Nonattrit *n*	4,345	4,125	4,045	3,973	3,896	3,807	3,083	1,665	1,077
	Attrit *n*	419	215	81	70	73	85	84	48	32
Hispanic	*d*	-0.31	-0.22	-0.25	-0.20	0.04	0.00	-0.09	-0.27	-0.74
	Nonattrit *n*	2,508	2,390	2,344	2,299	2,257	2,207	1,779	951	627
	Attrit *n*	175	112	46	45	42	48	38	36	18

Table 6.5. (continued)

		Attrition Period								
		3-Month	6-Month	9-Month	12-Month	15-Month	18-Month	21-Month	24-Month	27-Month
Amer. Ind.	*d*	-0.27	-0.18	-0.59	-0.24	0.73	-0.16	-0.40	0.73	--
	Nonattrit *n*	286	267	258	255	249	240	185	88	49
	Attrit *n*	36	19	9	3	5	10	6	7	1
Asian	*d*	-0.35	-0.48	-0.26	0.09	0.14	-0.56	0.12	0.04	-0.50
	Nonattrit *n*	712	690	676	667	657	648	523	262	166
	Attrit *n*	33	22	13	10	9	8	12	11	7
Other	*d*	-0.41	-0.40	-0.52	-0.34	-0.49	-0.41	0.25	-0.06	-1.15
	Nonattrit *n*	398	366	353	348	341	326	260	144	93
	Attrit *n*	58	32	9	7	7	15	12	9	5
Total	*d*	-0.40	-0.29	-0.23	-0.16	-0.12	-0.21	-0.28	-0.16	-0.24
	Nonattrit *n*	20135	18,854	18,354	17,973	17,566	17,094	13,553	6,902	4,411
	Attrit *n*	2360	1,234	482	367	403	450	455	286	151

Note. Effect sizes represent Attrit mean minus Nonattrit mean.

Table 6.6. Differential Validity Effect Sizes: Noncumulative Data for Air Force

		Attrition Period								
		3-Month	6-Month	9-Month	12-Month	15-Month	18-Month	21-Month	24-Month	27-Month
Gender										
Male	*d*	-0.55	-0.23	-0.12	-0.19	-0.05	-0.23	0.02	-0.15	-0.09
	Nonattrit *n*	13,148	12,889	12,748	12,674	12,610	12,540	12,453	12,354	12,286
	Attrit *n*	1,024	259	141	74	64	70	87	99	68
Female	*d*	-0.46	-0.09	0.02	-0.11	-0.29	-0.13	-0.12	0.12	0.14
	Nonattrit *n*	4,639	4,559	4,511	4,492	4,463	4,430	4,391	4,359	4,327
	Attrit *n*	490	80	48	19	29	33	39	32	32
Race										
White	*d*	-0.55	-0.24	-0.06	-0.11	-0.06	-0.34	-0.02	-0.14	-0.05
	Nonattrit *n*	10,300	10,082	9,959	9,901	9,845	9,783	9,707	9,621	9,562
	Attrit *n*	1,147	218	123	58	56	62	76	86	59
Black	*d*	-0.33	-0.17	-0.20	-0.08	-0.24	0.27	0.03	0.12	-0.11
	Nonattrit *n*	2,838	2,752	2,703	2,679	2,655	2,628	2,593	2,567	2,541
	Attrit *n*	249	86	49	24	24	27	35	26	26
Hispanic	*d*	-0.63	-0.05	-0.22	-0.09	0.00	0.02	-0.24	-0.24	0.35
	Nonattrit *n*	1,080	1,056	1,047	1,042	1,036	1,030	1,021	1,015	1,007
	Attrit *n*	61	24	9	5	6	6	9	6	8

Table 6.6. (continued)

		Attrition Period								
		3-Month	6-Month	9-Month	12-Month	15-Month	18-Month	21-Month	24-Month	27-Month
Other	*d*	-0.61	0.11	0.20	-1.15	-0.70	-0.96	-0.51	-0.55	0.01
	Nonattrit *n*	962	953	947	942	937	929	923	918	913
	Attrit *n*	51	9	6	5	5	8	6	5	5
Total	*d*	-0.53	-0.19	-0.08	-0.16	-0.13	-0.20	-0.03	-0.08	-0.03
	Nonattrit *n*	17,800	17,461	17,272	17,179	17,086	16,983	16,857	16,726	16,626
	Attrit *n*	1,514	339	189	93	93	103	126	131	100

Note. Effect sizes represent Attrit mean minus Nonattrit mean.

Table 6.7. Differential Validity Effect Sizes: Collapsed Noncumulative Data

		Army			Air Force		
		0-3 months	4-12 months	13-24 months	0-3 months	4-12 months	13-24 months
Gender							
Male	*d*	-0.41	-0.22	-0.14	-0.55	-0.15	-0.05
	Nonattrit *n*	17427	17468	17735	13148	13698	13852
	Attrit *n*	1716	1675	1408	1024	474	320
Female	*d*	-0.42	-0.14	0.02	-0.46	-0.01	-0.05
	Nonattrit *n*	2713	2949	2992	4639	4982	4996
	Attrit *n*	644	408	365	490	147	133
Race							
White	*d*	-0.40	-0.20	-0.09	-0.55	-0.11	-0.07
	Nonattrit *n*	11883	12132	12381	10300	11048	11167
	Attrit *n*	1639	1390	1141	1147	399	280
Black	*d*	-0.36	-0.13	-0.21	-0.33	-0.14	0.09
	Nonattrit *n*	4346	4399	4438	2838	2928	2975
	Attrit *n*	419	366	327	249	159	112
Hispanic	*d*	-0.31	-0.20	-0.04	-0.63	-0.06	-0.09
	Nonattrit *n*	2508	2480	2500	1080	1103	1114
	Attrit *n*	175	203	183	61	38	27

Note. Effect sizes represent Attrit mean minus Nonattrit mean.

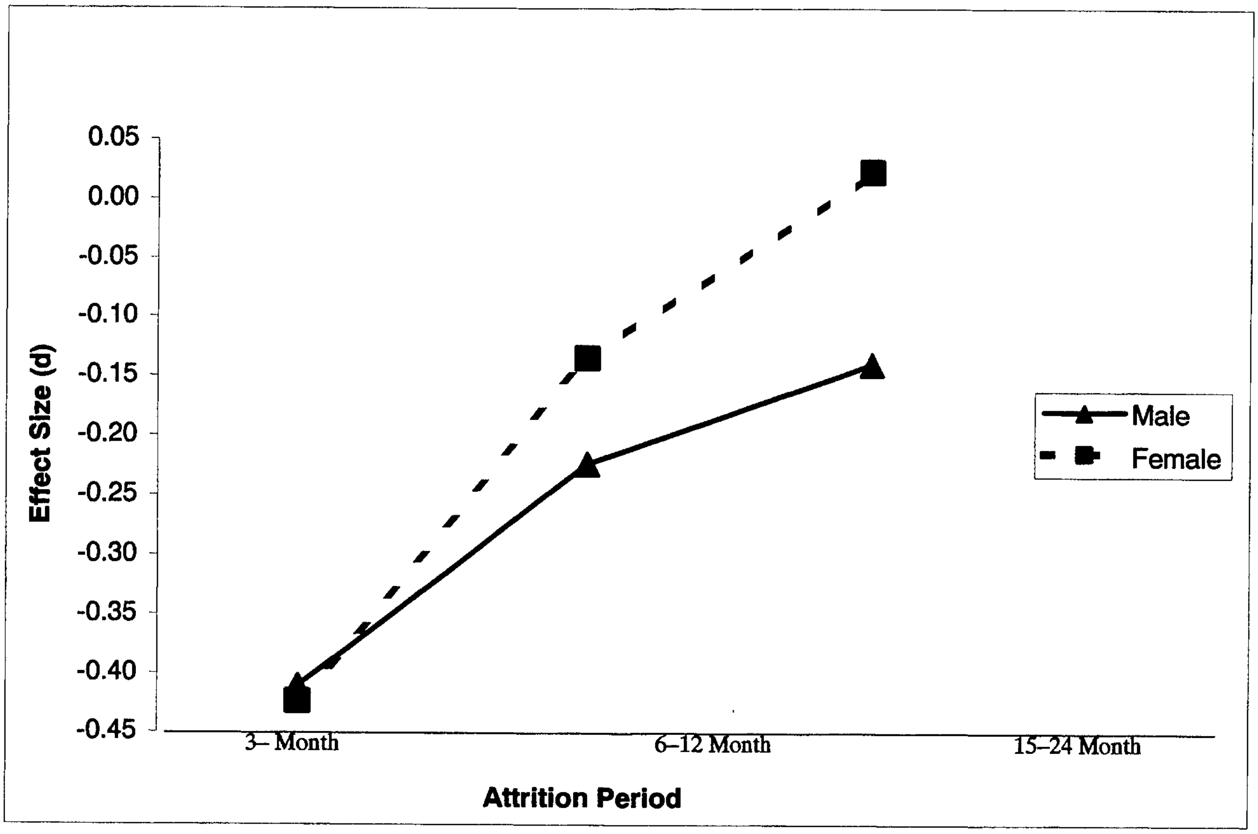


Figure 6.1. Differential validity by gender: Collapsed noncumulative data for Army.

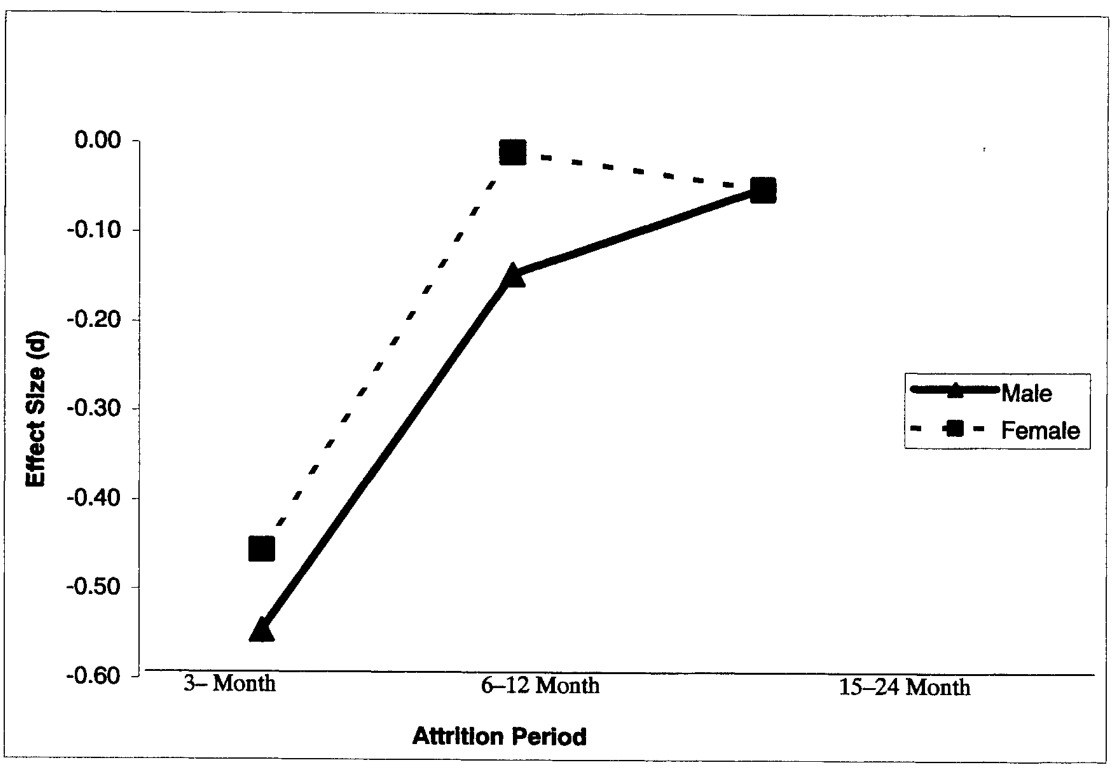


Figure 6.2. Differential validity by gender: Collapsed noncumulative data for Air Force.

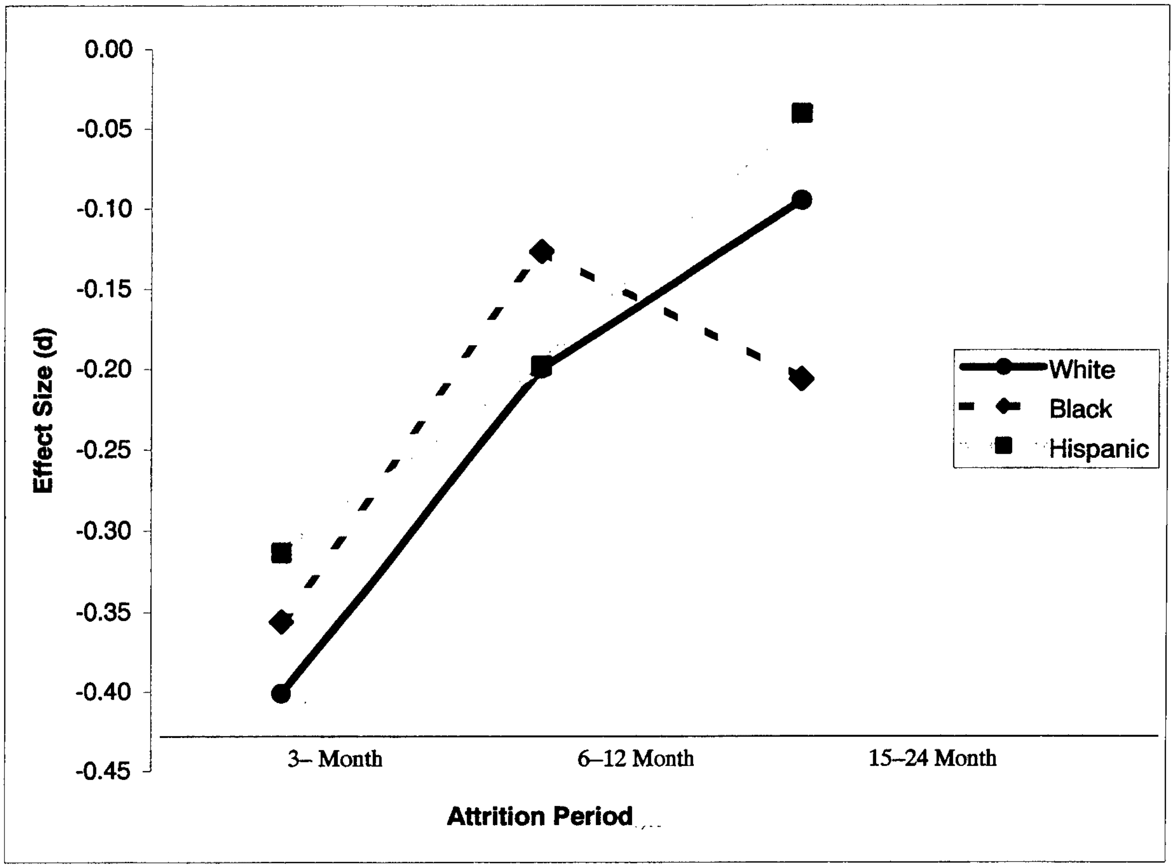


Figure 6.3. Differential validity by race: Collapsed cumulative data for Army.

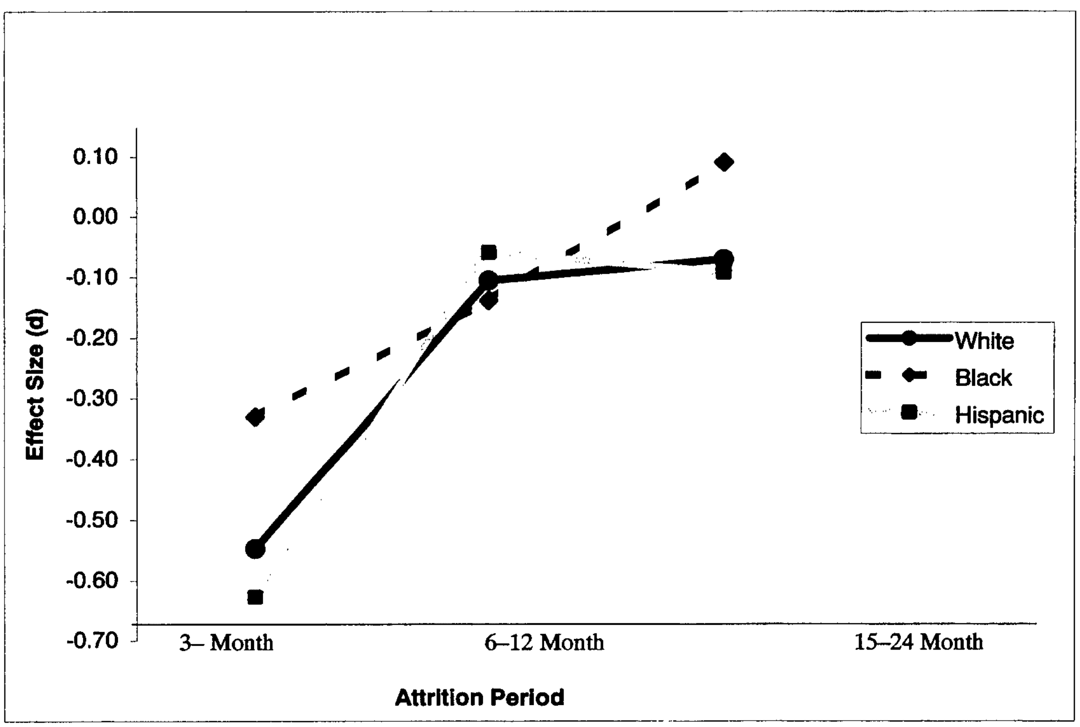


Figure 6.4. Differential validity by race: Collapsed cumulative data for Air Force.

An extensive literature exists on the investigation of differential prediction by race and gender in the cognitive ability domain (e.g., Dunbar & Novick, 1988; Houston & Novick, 1987; Hunter, Schmidt, & Rauschenberger, 1984; Schmidt, Pearlman, & Hunter, 1981; Valentine, 1977). Cognitive ability measures are among the most valid predictors of job performance across a wide variety of jobs (e.g., Hunter, 1986; Hunter & Hunter, 1984; Ree & Earles, 1991, 1992). However, Blacks and Hispanics on average score approximately 1 and 0.7 standard deviation points, respectively, below the White mean score on cognitive ability measures (Hunter, 1986).

These circumstances possibly explain the great interest in investigating predictive bias in the cognitive ability domain, especially with respect to race. The results of this research are summarized in the *SIOP Principles for the Validation and Use of Personnel Selection Procedures*: "There is little evidence to suggest that there is differential prediction for the sexes and the literature indicates that differential prediction on the basis of cognitive tests is not supported for the major ethnic groups (SIOP, 1987)."

While the SIOP *Principles* restrict their conclusion to the cognitive ability domain, we note that some researchers appear to assume that these findings generalize to other predictors. For example, Schmidt and Hunter (1998) state "For predictive fairness, the usual finding has been a lack of predictive bias for minorities and women.... .On some selection procedures (in particular, cognitive measures), subgroup differences on means are typically observed. On other selection measures (in particular, personality and integrity measures), subgroup differences are rare or nonexistent. For many purposes, the most relevant finding is the finding of lack of predictive bias. That is... the predictive interpretation of scores is the same in different subgroups (p. 272)."

Note the explicit discussion of both the ability and personality domains in the above statement. However, Sackett and Wilk (1994) reported that they could not locate any studies of differential prediction in the personality domain within the employment context. Their search included contacting the authors of two extensive meta-analytic reviews of validity of personality measures in the personnel selection context: Barrick and Mount (1991), and Tett, Jackson and Rothenstein (1991) reviewed 231 and 494 validity studies respectively, without being able to locate any studies of predictive bias. Saad and Sackett (2002) report an examination of predictive bias by gender for three personality characteristics, five criteria, and nine jobs using the Army Project A database (Campbell & Knapp, 2001). They report evidence of differential prediction in a number of instances, and it generally results in overprediction of female performance. This is consistent with findings in the cognitive ability domain that while differential prediction is rare, the limited instances where it is found result in overprediction of minority group performance.

In applying differential prediction analysis to the AIM-attrition relationship, one key feature is that the dependent variable – attrition – is dichotomous. The Ordinary Least Squares (OLS) regression model routinely used for differential prediction analysis is not ideal for use with a dichotomous criterion. Logistic regression is the approach of choice with a dichotomous criterion. While virtually all differential prediction research has used OLS and a continuous criterion, there is no conceptual reason why logistic regression cannot be used for the same purposes. The literature offers precedent for doing so. Rindskopf and Everson (1984), for example, used logistic regression to examine differential prediction by race in the college admissions process, where the dependent variable was dichotomous (admit/do not admit). Both

OLS and logistic regression analyses are reported here; they do not differ in the conclusions produced about differential prediction.

Appendix A (Tables A-1 and A-2) reports detailed results of OLS and logistic regression for the Army and Air Force data respectively for time periods up to 18 months. All individuals in the Army dataset had the potential for 18 months of service; this time period maximized the available sample size. Results generally show that the coefficients for the subgroup membership variables were significant and positive for gender and significant and negative for Blacks and Hispanics. The data indicate that attrition is underpredicted for females (women leave at a higher rate than predicted) and overpredicted for Blacks and Hispanics (Blacks and Hispanics leave at a lower rate than predicted). Note that the terms "overprediction" and "underprediction" have the opposite connotation from the usual application. In most settings the dependent variable is coded such that the desirable outcome (e.g., performance) takes on the larger values; in these data the undesirable outcome (attrition) is coded 1 while nonattrition is coded 0.

Differential prediction results are presented more intuitively in Tables 6.8 and 6.9 (OLS for Army and Air Force) and 8.10 and 8.11 (logistic regression for Army and Air Force), which apply the regression coefficients in the above analyses to generate a predicted attrition score for members of each subgroup at various AIM Adaptability Composite scores. Consider the logistic regression results in Table 6.10, at the current operational cutoff of 46. The overall logistic regression analysis on the entire sample shows that individuals with an AIM score of 46 have a 0.29 likelihood of attrition; thus the implicit intent of setting such a cutoff is to screen out anyone with a 0.29 or higher probability of attrition.

Looking first at the gender findings, one sees that at this cutoff women have a .46 likelihood of attrition: much higher than the target value of 0.29. Thus AIM underpredicts the likelihood that women will leave the service. The use of the AIM thus does not disadvantage women. While there is differential prediction by gender, it is not of the form that limits opportunities for women.

In contrast, consider the findings for Blacks and Hispanics. The cutoff of 46 corresponds to an attrition likelihood of 0.32 for Whites, 0.25 for Blacks, and 0.21 for Hispanics. Thus AIM overpredicts the likelihood of attrition for members of the Black and Hispanic groups. If an attrition likelihood of 0.29 is in fact the implicit target, the AIM cutoff that would be appropriate for Blacks would be 38, and the cutoff for Hispanics would be 28. Results for the Air Force parallel those for the Army.

Thus the regression analyses indicate that AIM overpredicts Black and Hispanic attrition. The use of the AIM for these subgroups would be viewed as unfair according to standard differential prediction methods. That is, there are statistically significant intercept differences between the majority and minority groups, which translate into the sizable differences noted in the above paragraph in the test cutoff scores that would result in a comparable attrition likelihood for Whites, Blacks, and Hispanics.

Table 6.8. Attrition Likelihood via OLS Regression: Army

		Attrition Likelihood by Subgroup								
Adaptability Score	% Passing	Overall (22,427)	Male (19,083)	Female (3,344)	White (13,463)	Black (4,758)	Hispanic (2,678)	Amer. Ind. (322)	Asian (745)	Other (454)
56	52.11	0.24	0.21	0.38	0.26	0.20	0.18	0.25	0.13	0.28
54	58.94	0.25	0.22	0.39	0.28	0.21	0.18	0.26	0.13	0.30
52	65.65	0.26	0.23	0.41	0.29	0.22	0.19	0.27	0.14	0.31
50	72.03	0.27	0.24	0.42	0.30	0.23	0.20	0.27	0.15	0.33
48	77.54	0.28	0.25	0.44	0.31	0.24	0.20	0.28	0.16	0.35
46	82.13	0.29	0.27	0.46	0.32	0.25	0.21	0.29	0.16	0.37
44	86.08	0.30	0.28	0.47	0.33	0.26	0.22	0.30	0.17	0.38
42	89.53	0.31	0.29	0.49	0.35	0.27	0.23	0.30	0.18	0.40
40	92.08	0.33	0.30	0.51	0.36	0.28	0.23	0.31	0.18	0.42
38	94.12	0.34	0.31	0.52	0.37	0.28	0.24	0.32	0.19	0.44
36	96.01	0.35	0.32	0.54	0.38	0.29	0.25	0.33	0.20	0.45
34	97.13	0.36	0.33	0.55	0.39	0.30	0.25	0.33	0.20	0.47
32	98.00	0.37	0.34	0.57	0.40	0.31	0.26	0.34	0.21	0.49
30	98.70	0.38	0.35	0.59	0.42	0.32	0.27	0.35	0.22	0.51
28	99.14	0.39	0.36	0.60	0.43	0.33	0.27	0.36	0.23	0.52
26	99.46	0.40	0.37	0.62	0.44	0.34	0.28	0.36	0.23	0.54
24	99.67	0.41	0.38	0.63	0.45	0.35	0.29	0.37	0.24	0.56

Note. Sample size listed in parentheses under subgroup.

Table 6.9. Attrition Likelihood via OLS Regression: Air Force

		Attrition Likelihood by Subgroup						
Adaptability Score	% Passing	Overall (16,443)	Male (12,204)	Female (4,239)	White (11,246)	Black (3,056)	Hispanic (1,133)	Other (1,008)
56	60.94	0.17	0.16	0.18	0.17	0.18	0.12	0.11
54	67.50	0.18	0.17	0.19	0.18	0.18	0.13	0.12
52	73.37	0.19	0.18	0.20	0.19	0.19	0.14	0.12
50	78.22	0.20	0.19	0.21	0.20	0.20	0.15	0.13
48	82.50	0.21	0.20	0.22	0.22	0.20	0.16	0.14
46	86.20	0.22	0.21	0.23	0.23	0.21	0.16	0.15
44	89.55	0.23	0.22	0.24	0.24	0.21	0.17	0.16
42	92.08	0.24	0.23	0.25	0.25	0.22	0.18	0.17
40	94.09	0.24	0.24	0.26	0.26	0.23	0.19	0.18
38	95.80	0.25	0.25	0.27	0.27	0.23	0.19	0.19
36	96.87	0.26	0.26	0.28	0.28	0.24	0.20	0.20
34	97.67	0.27	0.27	0.29	0.29	0.25	0.21	0.21
32	98.33	0.28	0.28	0.30	0.30	0.25	0.22	0.22
30	98.80	0.29	0.29	0.31	0.31	0.26	0.23	0.23
28	99.27	0.30	0.30	0.31	0.32	0.26	0.23	0.24
26	99.51	0.31	0.31	0.32	0.33	0.27	0.24	0.25
24	99.71	0.32	0.31	0.33	0.34	0.28	0.25	0.26

Note. Sample size listed in parentheses under subgroup.

Table 6.10. Attrition Likelihood via Logistic Regression: Army

		Attrition Likelihood by Subgroup								
Adaptability Score	% Passing	Overall (22,427)	Male (19,083)	Female (3,344)	White (13,463)	Black (4,758)	Hispanic (2,678)	Amer. Ind. (322)	Asian (745)	Other (454)
56	52.11	0.23	0.21	0.37	0.26	0.20	0.18	0.25	0.12	0.27
54	58.94	0.24	0.22	0.39	0.27	0.21	0.18	0.26	0.13	0.29
52	65.65	0.25	0.23	0.41	0.28	0.22	0.19	0.26	0.14	0.31
50	72.03	0.27	0.24	0.42	0.29	0.23	0.20	0.27	0.14	0.32
48	77.54	0.28	0.25	0.44	0.31	0.24	0.20	0.28	0.15	0.34
46	82.13	0.29	0.26	0.46	0.32	0.25	0.21	0.29	0.16	0.36
44	86.08	0.30	0.28	0.47	0.33	0.26	0.22	0.30	0.17	0.38
42	89.53	0.32	0.29	0.49	0.35	0.27	0.23	0.30	0.18	0.41
40	92.08	0.33	0.30	0.51	0.36	0.28	0.24	0.31	0.19	0.43
38	94.12	0.34	0.31	0.53	0.37	0.29	0.24	0.32	0.20	0.45
36	96.01	0.36	0.33	0.54	0.39	0.30	0.25	0.33	0.20	0.47
34	97.13	0.37	0.34	0.56	0.40	0.32	0.26	0.34	0.21	0.49
32	98.00	0.38	0.35	0.58	0.42	0.33	0.27	0.35	0.23	0.51
30	98.70	0.40	0.37	0.59	0.43	0.34	0.28	0.36	0.24	0.53
28	99.14	0.41	0.38	0.61	0.45	0.35	0.29	0.37	0.25	0.56
26	99.46	0.43	0.40	0.63	0.46	0.37	0.30	0.37	0.26	0.58
24	99.67	0.44	0.41	0.64	0.48	0.38	0.31	0.38	0.27	0.60

Note. Sample size listed in parentheses under subgroup.

Table 6.11. Attrition Likelihood via Logistic Regression: Air Force

		Attrition Likelihood by Subgroup						
Adaptability Score	% Passing	Overall (16,443)	Male (12,204)	Female (4,239)	White (11,246)	Black (3,056)	Hispanic (1,133)	Other (1,008)
56	60.94	0.16	0.15	0.18	0.17	0.18	0.12	0.10
54	67.50	0.17	0.16	0.19	0.18	0.18	0.13	0.10
52	73.37	0.18	0.17	0.20	0.19	0.19	0.14	0.12
50	78.22	0.19	0.18	0.21	0.20	0.20	0.14	0.13
48	82.50	0.20	0.20	0.22	0.21	0.20	0.15	0.14
46	86.20	0.21	0.21	0.23	0.22	0.21	0.16	0.15
44	89.55	0.23	0.22	0.24	0.24	0.22	0.17	0.17
42	92.08	0.24	0.23	0.25	0.25	0.22	0.18	0.18
40	94.09	0.25	0.25	0.26	0.26	0.23	0.20	0.20
38	95.80	0.26	0.26	0.27	0.28	0.24	0.21	0.21
36	96.87	0.28	0.27	0.29	0.29	0.25	0.22	0.23
34	97.67	0.29	0.29	0.30	0.31	0.25	0.23	0.25
32	98.33	0.31	0.30	0.31	0.32	0.26	0.25	0.27
30	98.80	0.32	0.32	0.33	0.34	0.27	0.26	0.29
28	99.27	0.34	0.33	0.34	0.36	0.28	0.28	0.31
26	99.51	0.35	0.35	0.35	0.37	0.29	0.29	0.34
24	99.71	0.37	0.37	0.37	0.39	0.29	0.31	0.36

Note. Sample size listed in parentheses under subgroup.

Revisiting Differential Prediction Methodology: The Omitted Variables Problem

These findings are quite troubling, and resulted in a critical reexamination of the analytic approach used for differential prediction analysis. This reanalysis led to a focus on an underexamined issue in differential prediction analysis, namely, the omitted variables problem.

Attempts to estimate regression coefficients rely on a set of fundamental assumptions. The key assumption here is that of a fully specified model. In other words, the assumption is that all determinants of the dependent variable are included in the model. Failure to include all determinants results in a mis-specified model, a result often referred to as the "omitted variables problem."

An omitted variable is only a problem under a specific set of circumstances. If a variable that is related to the dependent variable but uncorrelated with any measured independent variable is omitted, the result is a model that fits more poorly and has a larger error term. The regression coefficients for the measured independent variables, however, are not biased by the omission of such a variable. In contrast, if a variable that is related to the dependent variable and correlated with a measured independent variable is omitted, the regression coefficient for the measured independent variable can be biased.

This fact is well known. However, it is rare for there to be a criterion variable that is so well understood that all relevant independent variables are known in advance and thus included in the model. Thus potential omitted variables are commonly a problem in psychological research, perhaps to the point that they become background noise; the possibility is always present, and so is rarely attended to.

That omitted variables may be an issue in differential prediction analysis was recognized early: Within 3 years of the Cleary (1968) formulation, Linn and Werts (1971) called attention to the omitted variables problem in using regression analysis for this purpose. Differential prediction analysis interprets the coefficient for the subgroup variable as indicating intercept differences by subgroup and interprets the coefficient for the interaction between subgroup membership and the test in question as indicating slope difference by subgroup. If a variable correlated with the criterion is also correlated with either subgroup membership or with the test in question, the coefficients for subgroup membership and for the subgroup/test interaction may be biased. As will be shown below, this bias can lead to a conclusion that differential prediction is present when, in fact, it is not. It can also lead to a conclusion of no differential prediction when, in fact, differential prediction is present.

To demonstrate this, we generated a dataset with 1,000 cases. We generated three random variables, each in standard score form: A, B, and E. View A as a cognitive ability test, B as a measure of conscientiousness, and E as random error. As random variables these three factors are, by definition, uncorrelated, a finding that corresponds to those in the research literature regarding the relationship between ability and conscientiousness. Assume that the sample is made up of a Black subgroup (n=100) and a White subgroup (n=900). Mimicking common findings in the literature, we lowered the ability test scores by one standard deviation for each member of the Black subgroup. Members of the Black and White subgroups did not differ on the conscientiousness scale.

We then created a criterion variable (i.e., job performance) as A+B+2E. In other words, we created a simple system in which performance is solely a function of A, B, and random error. By definition A and B are unbiased predictors of this criterion.

Now imagine that a researcher is interested in determining whether the cognitive ability measure exhibits differential prediction by race. Imagine also that the researcher has not measured conscientiousness. We know the true state of affairs (i.e., the relationship between conscientiousness and performance), but let us assume that the researcher does not. The researcher would run a regression model entering cognitive ability, group membership, and the interaction between the two. As the results in Table 6.12 show, ability is related to performance, with no slope or intercept differences by race. This is the correct finding; ability is by definition in this simulation an unbiased predictor of performance. The omission of the conscientiousness variable does not cause problems because, while correlated with performance, conscientiousness is not correlated with either race or ability, and thus the coefficients for race and ability are unbiased.

Table 6.12. Regression Model Entering Cognitive Ability, Racial Group Membership, and the Interaction Between the Two

	B	Std. Error	*p*
(constant)	-.433	.343	.208
Ability	.672	.242	.005
Race	.423	.352	.230
Race X Ability	.374	.254	.141

Now imagine that a researcher is interested in determining whether the conscientiousness measure exhibits differential prediction by race. Imagine also that the researcher has not measured ability. We know the true state of affairs (i.e., the relationship between ability and performance), but let us assume that the researcher does not. The researcher would run a regression model entering conscientiousness, group membership, and the interaction between the two. The results are shown in Table 6.13.

Table 6.13. Regression Model Entering Conscientiousness, Racial Group Membership, and the Interaction Between the Two

	B	Std. Error	*p*
(constant)	-1.114	.227	.000
Conscientiousness	.876	.262	.001
Race	1.185	.239	.000
Race X Conscientiousness	.207	.272	.446

Here we see that an erroneous conclusion is reached: The coefficient for race is significant, suggesting intercept differences between the groups. But the simulated data were designed such that conscientiousness is an unbiased predictor. Here is an example of the omitted variable problem. Ability is related to performance, but also correlated with race. With ability omitted from the equation, the shared variance between ability and race is attributed to race, biasing the race coefficient and leading to the erroneous conclusion of differential prediction for the conscientiousness variable.

What is the solution? Clearly, the omitted variable(s) must be identified, measured, and included in the model. Table 6.14 presents the results of an alternative regression that included ability as a control variable in addition to conscientiousness, race, and the race X conscientiousness interaction.

Thus when the omitted variable – ability – is in the equation, the race variable is not significant. The correct findings are produced: Both ability and conscientiousness are revealed as predictive of performance, with no differential prediction.

Table 6.14. Regression Model Entering Conscientiousness, Racial Group Membership, Ability, and the Interaction Between Race and Conscientiousness.

	B	Std. Error	*p*
(constant)	.012	.212	.955
Conscientiousness	.1.087	.232	.000
Race	.031	.242	.888
Ability	1.060	.064	.000
Race X Conscientiousness	.028	.241	.907

This demonstration shows how an omitted variable can lead to the illusion of differential prediction when none is in reality present. We turn now to the converse, namely the situation in which differential prediction is, in fact, present. To do this, we created bias in the conscientiousness variable by lowering the scores of the Black sample by one standard deviation. We did not similarly adjust criterion scores. Thus we created the definitional characteristic of bias, namely, predictor score differences between subgroups that do not correspond to criterion differences.

We then performed traditional differential prediction analysis on the conscientiousness variable, treating ability as an omitted variable. This produced the results shown in Table 6.15. Thus conscientiousness, which is now a truly biased predictor in this simulation, emerges as unbiased. The omission of ability biases the coefficients of variables correlated with ability (e.g., race).

Table 6.15. Regression Model, Treating Ability as an Omitted Variable, Entering Conscientiousness and Racial Group Membership, and the Interaction Between Race and Conscientiousness.

	B	Std. Error	*p*
(constant)	-.237	.353	.501
Conscientiousness	.876	.262	.001
Race	.308	.361	.393
Race X Conscientiousness	.207	.272	.446

Table 6.16 presents the results when ability is added to the model. Race is now significant, with the performance of Blacks underpredicted. The bias we built into the conscientiousness variable is now revealed in the differential prediction analysis, once the omitted variable problem is removed. In sum, we have shown that omitted variables can cause unbiased predictors to appear biased and biased predictors to appear unbiased.

Table 6.16. Regression Model, Treating Ability as an Omitted Variable, Entering Conscientiousness, Racial Group Membership, Ability, and the Interaction Between Race and Conscientiousness.

	B	Std. Error	*p*
(constant)	1.099	.323	.001
Conscientiousness	1.087	.232	.000
Race	-1.055	.330	.001
Ability	1.060	.064	.000
Race X Conscientiousness	.028	.241	.907

Omitted Variables and the Analysis of AIM

In light of this analysis, there are now two competing hypotheses: (a) AIM is biased against Black and Hispanic applicants, as it overpredicts attrition for these groups; and (b) the appearance of overprediction is an artifact resulting from the omission of one or more variables from the regression models used to examine differential prediction.

This suggests that the first line of inquiry should be the examination of possible omitted variables. If a set of omitted variables could be identified and measured, and if including these variables in the model eliminated the differential prediction on the basis of race, the hypothesis that AIM is biased would be refuted, assuming that the model were now fully specified (i.e., all omitted variables were now included).

It is useful to consider the needed characteristics of omitted variables that, if included, would reduce or eliminate the race effect. As Blacks and Hispanics are less likely to leave the service, an omitted variable would reduce the race effect if it correlated positively with membership in Black or Hispanic groups and correlated negatively with attrition, or the converse of this (i.e., correlated negatively with membership in Black or Hispanic groups and correlated positively with attrition). For example, consider the Armed Forces Qualification Test (AFQT). It correlates negatively with membership in Black and Hispanic groups (i.e., Blacks and Hispanics have lower mean AFQT scores) and also correlates negatively with attrition (i.e., individuals with lower AFQT scores are more likely to leave the service). Thus AFQT is not a variable that, if included in the model, would reduce the apparent differential prediction. As a variable correlated with both race and attrition, though, it would be a necessary variable to include in a fully specified model.

Thus to eliminate the observed differential prediction by race, omitted variables must fully account for the apparent race-attrition relationship. Theoretical work to develop hypotheses about reasons for lower attrition in Black and Hispanic subgroups is needed, followed by data collection to test these hypothesized relationships. Factors worthy of exploration might include socioeconomic status, perceived alternative labor market opportunities, and cultural values regarding completing a tour of duty.

The Omitted Variables Problem in the Professional Literature

Given the potential for omitted variables to influence conclusions about differential validity, and given that this issue has been raised as early as 1971, one might think that omitted variables would be routinely considered in applications of differential prediction analysis. This, however, is not true. As a way of gauging the degree to which attention was paid to the omitted variables problem, we drew a sample of 33 published applications of differential prediction analysis from the published literature in education and psychology. While the omitted variables issue was mentioned in 6 of the 33 studies, only 2 of the 33 studies actually included variables beyond test, subgroup, and the test/subgroup interaction in their analysis. Thus attention to this issue is a rare exception, rather than the norm.

Why has this issue been generally disregarded? We speculate that one reason is that differential prediction analysis has been applied almost exclusively to cognitive tests, and the typical finding with these tests is that, if differential prediction is found, it is in the form of overprediction of racial subgroup performance. Thus racial minority groups are not harmed by the differential prediction, and thus the prediction system is not carefully scrutinized further. We suggest that if underprediction were found, much more attention would focus on the omitted variables problem.

As a concrete example, there is one domain where underprediction of subgroup performance is found for a protected subgroup, namely, the use of the SAT in predicting college performance. The performance of women is commonly underpredicted, though not by a large amount. Stricker, Rock, and Burton (1993) examined a large set of potential omitted variables to gain insight into this underprediction. Several key variables were found, the inclusion of which reduced the underprediction to a large degree. Key variables included number of hours studying and the percentage of readings and other assignments completed. With SAT held constant, women studied more and completed more of the readings and other assignments. Without these variables in the model, variance in academic performance due to these variables is erroneously attributed to gender, resulting in differential prediction.

It is also worth noting that the omitted variables problem is not mentioned in the *AERA/APA/NCME Standards for Educational and Psychological Testing* (1999) or the SIOP *Principles for the Validation and Use of Personnel Selection Procedures* (1987). Both acknowledge differential prediction analysis as the appropriate means of detecting predictive bias. In light of the issues raised here, it is not clear that the confidence in the approach is well founded.

Differentiating Between Bias in a Test and Bias in a Selection System

What if differential prediction disappears when an omitted variable is added to the equation? This question highlights a crucial distinction between bias in a test and bias in a selection system. If it can be shown that differential prediction is present when a test is examined alone in the traditional moderated regression framework, but disappears once all relevant variables are included in the regression equation, one can conclude that the test itself is not biased. One also concludes that a selection system that includes all of the variables in the fully specified model is not biased. However, the decision to use the test alone as the selection system, without also including the other variables in the fully specified model, would result in a biased selection system.

For example, consider the Stricker et al. (1993) study predicting academic performance discussed above. The academic performance of women was underpredicted when the SAT alone was used as the selection system. Adding other variables (e.g., study habits) resulted in a more fully specified equation in which the underprediction essentially disappeared. This leads to the conclusion that the SAT itself, as a test, is not biased against women. But the use of the SAT alone as the basis for selection would be biased, as such a selection system would indeed underpredict women's performance. Only if a composite of the SAT and the relevant other variables (e.g., study habits) were used as the basis for selection would the selection system be unbiased.

In short, if conclusions about differential prediction vary depending on whether variables other than the focal test are included in the regression equation, then those variables must be included in the selection system for the selection system to be unbiased.

A question of considerable interest is what to do if the relevant omitted variables either have not been identified or are not measured? Sackett and Wilk (1994) argue that group-specific regression equations are appropriate. The *Standards* make the same argument. Conceptually, one is computing a predicted criterion for each person, and is using all information available to make this prediction as accurate as possible. Sackett and Wilk argue that this could be interpreted as permissible under the Civil Rights Act of 1991, since the prohibition in that act was against within-group scoring and other score adjustment techniques designed to increase minority representation. There was no argument by proponents of within-group scoring that such scores were more accurate – it was a representation argument, not a validity maximization argument. Computing predicted performance scores using separate regressions is designed to maximize validity, not to maximize representation.

The problem, though, is that one would only make this validity maximization argument in settings where minority performance is underpredicted. One would not see the same argument applied in the far more common setting in which minority performance is overpredicted. If one were serious about validity maximization, one would use separate regressions in that situation as well, resulting in the reduction of minority representation relative to the use of a common regression line. The fact that one appeals to validity maximization only when separate regression lines would increase minority representation argues that minority representation is, in fact, the motivation. Therefore, the use of separate regression lines in situations where minority performance is underpredicted arguably falls within the purview of the ban on score adjustment in the Civil Rights Act of 1991.

Legal and Professional Obligations to Conduct Differential Prediction Analysis

A key question is: If there is no adverse impact, is there an obligation to conduct differential prediction analysis? Legally, the answer is no. Under the *Uniform Guidelines on Employee Selection Procedures*, all legal scrutiny of a selection system hinges on a finding of adverse impact. As the AIM Adaptability Composite score shows very small subgroup differences, and hence no adverse impact, there is no legal obligation to conduct validity or differential prediction analyses.

Professionally, the SIOP *Principles* discuss differential prediction, but do not address the issue of what motivates conducting such an analysis. The *APA Standards* also discuss differential prediction, stating that differential prediction is the source of evidence for bias or lack of bias. As the *Standards* reject adverse impact as the basis for conclusions about bias, one might, by inference, infer that the presence or absence of adverse impact is not determinative as to whether differential prediction analysis should or should not be done. Conceptually, since differential prediction can exist regardless of the presence or absence of adverse impact, one might argue that a lack of adverse impact should not lead to the conclusion that differential prediction analysis is not necessary. But the *Standards* do not address directly the circumstances that would motivate conducting differential prediction analysis.

Conclusions About Differential Prediction

The following points summarize the current state of affairs regarding the AIM Adaptability Composite score:

- The score exhibits differential prediction against Blacks and Hispanics: Members of these subgroups are less likely to leave the service than AIM would predict. The use of a common regression line thus shows predictive bias against members of these groups.

- It is not clear whether AIM is truly biased, or whether the obtained results reflect an omitted variables problem in which the effects of omitted variables correlated with race (e.g., socioeconomic status, perceived labor market alternatives, cultural values regarding completing a tour of duty) are attributed, falsely, to race.

- The data available for examination here do not include likely omitted variables.

- If key omitted variables were identified, and if their inclusion in the model eliminated the race effect, one would conclude that the AIM itself did not have predictive bias. But these omitted variables would have to be included along with the AIM in a selection system in order for the selection system to be free from predictive bias.

- For employers covered by the Civil Rights Act of 1964, and to whom the 1978 *Uniform Guidelines on Employee Selection Systems* apply, there is no legal obligation to conduct, report, or act on the results of differential prediction analysis in the absence of adverse impact; no adverse impact results from the use of the AIM.

- Professional standards do not directly address the circumstances under which one has an obligation to conduct and report differential prediction analyses.

- If one wished to use AIM until further research to identify key omitted variables was conducted, one possibility is the use of separate regression equations for each subgroup, a practice that would eliminate the differential prediction. It is unclear whether such a practice would be permissible for organizations covered by the Civil Rights Act of 1991.

CHAPTER 7: ALTERNATIVE METHODOLOGIES FOR PREDICTING ATTRITION IN THE ARMY: THE NEW AIM SCALES

Fritz Drasgow, Wayne C. Lee, Steve Stark, and Oleksandr S. Chernyshenko
University of Illinois at Champaign-Urbana

Prior to our Pre-Implementation Research Program (1998 – 1999), evaluations of AIM's psychometric properties were limited by the constraints of relatively small sample sizes. The authors of this chapter report their assessment of AIM's psychometric performance, based on classical test statistics, using a large sample (n = 22,000) from this program's research database.

ARI developed the AIM Adaptability Composite for predicting first-term attrition. This composite was the primary focus of investigation in our large research sample, and later was implemented as the operational attrition screen under the GED Plus Program. We felt it was important to explore alternative approaches to creating an AIM attrition composite. Perhaps there might be a more optimal method for scoring AIM and/or weighting the components of such a composite. This was the issue addressed in the effort reported here. The findings indicated that there is potential for improving AIM's validity through using alternative scoring approaches. We have continued exploring these options well beyond the timeframe of the efforts presented in this report. However, unlike the work reported here, our more recent efforts have used the operational database.

Introduction

The purpose of the analyses described in this chapter was to examine alternative methods of predicting attrition using the AIM scales. We used the 22,666 Regular Army cases contained in the Army AIM Grand Research Database. The analyses included:

- basic analyses (classical test statistics, including item-total correlations and coefficient alpha at the scale level; principal components analysis; and correlations between each stem and each scale with retention);
- multiple linear and logistic regression of retention onto the AIM scales;
- classification tree modeling of attrition with the AIM scales at 12 months; and
- item response theory (IRT) modeling of the polytomously scored AIM scales and optimal classification via the Neyman-Pearson lemma and the IRT models applied to the scales.

Before any analyses were carried out, 22 individuals were excluded from the database because their Army component identifier indicated that they were members of the Army National Guard or the Army Reserve (i.e., not Regular Army), or information on their Army component was not available.

Basic Analyses

Classical Test Statistics

Classical test statistics were computed for all of the items of the six AIM content scales: Scales A through F. The analyses were conducted using the AIM trichotomous item scoring (2, 1, and 0). If a stem had a negative corrected item-total correlation, it was removed from the scale and the statistics were recomputed. This was the case with two stems, the first stem in both Scales B and D. These stems were also removed for the IRT analyses. Coefficient alpha varied between the scales from .70 to .57. These low to moderate values for coefficient alpha may be due to multidimensionality (see below) or the small number of stems comprising some of the scales. Compared to the values found with the pretest data (scales administered one-by-one with Likert-type ratings), these values are indeed lower (Heggestad, Lightfoot, & Waters, 1999, p. 24). It may be the case that these findings were due to the partially-ipsative nature of these items as compared to the Likert-type presentation with the pretest respondents. Appendix B (Table B.1) provides an example of these results for Scale C.

Principal Components Analysis

To determine the dimensionality of each scale at the stem level, principal-components analyses were carried out on the stems of each of the six content scales. The results show scant evidence of multidimensionality within each scale. In examining the eigenvalues for each of these scales, a sharp elbow is apparent in each case. Figure 7.1 clearly shows the sharp elbow for each scale.

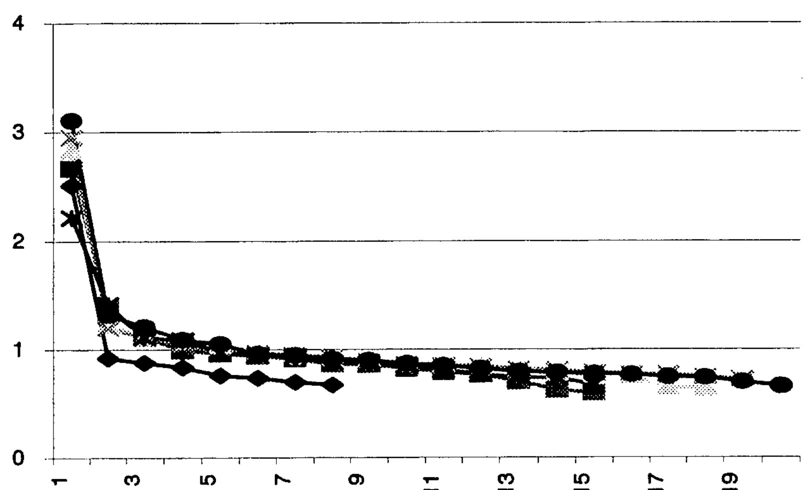


Figure 7.1. Scree plot following principal components analysis for the six content scales.

Stem-Retention Correlations

To determine whether the individual stems were related to retention, each stem within each item was correlated with retention. As described in Chapter 1, each tetrad comprises two positively worded and two negatively worded trait stems. Trait stems were trichotomously scored (0, 1, or 2) and validity stems were dichotomously scored (0, 1). The *retention* variable (0 indicating a leaver, 1 indicating a stayer) was created for various time intervals (i.e., 3, 6, 9, 12, 15, 18, and 21 months). The findings suggest that many stems are positively related to retention, as should be the case. A few stems had low negative or near-zero correlations, indicating that the stem does not predict retention, but may be useful in another context (i.e., other stems in an item's tetrad may have large correlations). As an example, the point-biserial correlations for Scale C are provided in Table 7.1. A full set of results will be available from ARI on a restricted basis.

The same analyses were carried out at the scale level across all seven scales in the AIM. These results are summarized in Table 7.2.

Table 7.1. Stem-Retention Correlations for Scale C

Retention at specified time interval (months)	Stem 1	2	3	4	5	6	7	8
3	0.076	0.053	0.082	0.053	0.071	0.060	0.073	0.076
N	22,565	22,540	22,539	22,545	22,594	22,587	22,580	22,493
6	0.086	0.066	0.090	0.058	0.083	0.067	0.085	0.090
N	22,565	22,540	22,539	22,545	22,594	22,587	22,580	22,493
9	0.085	0.068	0.094	0.060	0.083	0.069	0.089	0.092
N	22,565	22,540	22,539	22,545	22,594	22,587	22,580	22,493
12	0.086	0.070	0.097	0.060	0.080	0.068	0.087	0.091
N	22,565	22,540	22,539	22,545	22,594	22,587	22,580	22,493
15	0.089	0.079	0.102	0.067	0.083	0.072	0.090	0.097
N	19,574	19,547	19,551	19,559	19,603	19,595	19,597	19,526
18	0.104	0.095	0.111	0.087	0.101	0.086	0.104	0.135
N	13,096	13,086	13,092	13,096	13,124	13,118	13,116	13,074
21	0.106	0.094	0.110	0.090	0.105	0.099	0.111	0.146
N	10,401	10,389	10,395	10,402	10,425	10,419	10,421	10,379

Note. All correlations significant at $p<.01$.

Table 7.2. Scale-Retention Correlations

Retention at specified time interval (months)	Scale						
	A	B	C	D	E	F	G
3	0.114	0.065	0.122	0.076	0.004 ns	0.071	0.024
6	0.122	0.068	0.140	0.075	0.008 ns	0.082	0.027
9	0.119	0.068	0.143	0.075	0.002 ns	0.080	0.023
12	0.116	0.069	0.142	0.074	-0.001 ns	0.077	0.019
15	0.117	0.081	0.151	0.076	-0.002 ns	0.074	0.014 ns
18	0.131	0.115	0.181	0.082	0.006 ns	0.084	-0.006 ns
21	0.139	0.142	0.188	0.091	0.012 ns	0.091	-0.008 ns

Note. Unless indicated, correlations are significant at $p < .01$.

Regression Analyses

Basic Linear Regression

Basic regression of retention onto the AIM scales was carried out with the AIM scales separately, with their squares and cubes (to examine nonlinearities), multiple regression with all the AIM scales, and finally multiple regression with control variables (dummy variables for AFQT, number of dependents, gender, and race). The time interval selected was 12 months. For these analyses, we began by dividing the sample into halves by selecting the odd-numbered cases and the even-numbered cases. Regression equations obtained from the odd-numbered cases were compared to the regression coefficients estimated from the other half-sample. Very similar regression coefficients were obtained across the two half-samples, which should not be surprising because each half-sample contained more than 10,000 cases. Because these initial analyses indicated that the sample was large enough to ensure cross-validity, the regression equations were re-estimated using the entire data set.

The regression coefficients of the squared terms were usually negative and the increases in R^2 obtained by adding these terms were small. The regression coefficients of the squared terms were often significant, but we interpret this as an artifact of multicollinearity between the linear and squared terms. The Adaptability Composite correlated .132 with 12-month retention. The regression of retention on all seven AIM scales produced a multiple correlation of .158. Thus, a modest improvement in prediction accuracy was obtained by using ordinary least squares regression weights. It was clear from the results that most square and cube terms played a small role in the regression equations.

Logistic Regression

The assumptions necessary for linear regression are not met by the 12-month retention variable (coded as 0, 1); thus, we conducted a logistic regression of retention onto the AIM scales.

Again, we split the sample into halves (odd- and even-numbered cases). Logistic regression coefficients were estimated from each half-sample; results across the samples were very similar.

Logistic regression equations were examined for the AIM scales (a) separately, (b) with their squares and cubes (to examine nonlinearity), (c) with all the AIM scales, and (d) with control variables. Scales A and C are important predictors of retention. When other control variables are entered into the logistic regression equation (e.g., AFQT, gender, race), Scale F receives a significant regression coefficient in addition to Scales A and C.

Classification and Regression Trees

Description of CART

Another method of predicting retention can be found through the application of classification and regression trees (CART) (Breiman, Freidman, Olshen, & Stone, 1997). CART can be used to predict any number of outcomes based on a set of predictor variables. The basis of CART is a "decision tree" that is grown with branches stemming from a set of nodes based on binary splits (answers to "yes/no" questions) and ending with terminal nodes. These terminal nodes correspond to the categories of the outcome variables. With CART, separate "paths" can be delineated that lead to the same outcome.

To assess classification accuracy, CART utilizes "v-fold cross validation." In this procedure, the sample is divided into v subsamples (in our analyses, we used $v = 10$). Then CART grows a decision tree after combining v-1 subsamples and assesses the classification accuracy using the hold-out subsample. This process is iterated so that v-1 subsamples are combined and used to grow decision trees and each of the v samples is used as the hold-out sample once. Classification accuracy is estimated as the average classification accuracy across the v holdout subsamples.

Input

For this analysis, we used the CART 4.0 statistical software package (Breiman et al., 1997). A data set with 22,328 enlisted personnel was created containing all seven AIM scales and the retention variable to 12 months. The 12-month time interval was selected because this provided the CART program with a sufficient number of respondents with which to "grow" a tree. The CART program exhaustively examines all possible binary splits with each predictor and arranges them into separate trees with the best predictor being the "root" of the trees. CART then chooses the best tree based on the costs (misclassification rates) associated with the outcome variable. Furthermore, CART selects trees so as to minimize the number of terminal nodes that consist of only a handful of cases. CART also tests the trees grown against a holdout sample.

Results

The analysis yielded 39 trees ranging in complexity from two terminal nodes to a tree with 2,802 terminal nodes and a depth of 51 levels or tiers. However, the larger trees exhibited high rates of misclassification among soldiers who stayed in the Army (up to 60%). Of particular interest were five trees resulting from this analysis that had relatively low rates (31% to 33%) of misclassification of soldiers who stayed in the Army. Table 7.3 summarizes the misclassification rates for these five trees estimated by v-fold cross-validation. Note that more complex decision trees did not improve classification accuracy and, in some cases, provided much worse results upon cross-validation.

Table 7.3. Misclassification Rates for Five Classification Trees

Number of terminal nodes	"False positives" (misclassification of nonattritees)	"Hits" (correct classification of attritees)
3	31.14%	45.32%
6	34.23%	48.19%
7	33.64%	47.40%
11	33.40%	47.13%
18	32.09%	45.68%

For purposes of illustration, the first and third of these trees are depicted in Figures 7.2 and 7.3 respectively (left branches indicate a "yes" response to the parent node; right branches indicate a "no" response). For example, Figure 7.2 shows that the root (i.e., initial) node splits the sample on the basis of Scale C scores; individuals with "C-scores" less than or equal to 8.5 are predicted to be attritees, and individuals with scores greater than 8.5 are branched to another node. In this node, individuals with relatively high C-scores (i.e., greater than 8.5) but low A-scores (less than or equal to 14.89) are predicted to attrit. Only individuals with high C- scores and high A-scores are predicted to be nonattritees.

CART also rank-orders the relative importance of the predictor variables. In this analysis, CART identified the two best predictors as Scales C and A. Scales F, D, and B played a smaller role in these classification trees, whereas Scales E and G played nearly insignificant roles.

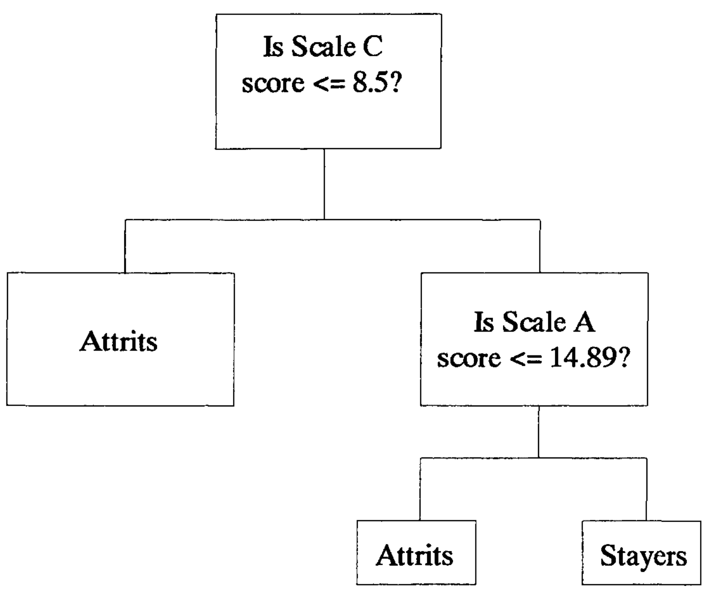


Figure 7.2. Classification tree with 3 terminal nodes.

AIM Prediction of Attrition Using Item Response Theory

Summary of Procedures

Item response theory comprises a set of psychometric models for describing how individuals respond to items. These models contain a person parameter, often denoted by the Greek letter theta (θ), that represents an individual's standing on the trait assessed by a test or scale. Each item is characterized by one or more item parameters. The two-parameter logistic model, for example, includes one item parameter that characterizes item difficulty (sometimes referred to as item extremity in personality assessment) and a second parameter that characterizes item discrimination. Samejima's (1969) graded response model can be used to model ordered categorical responses; it includes one parameter that represents the item's discrimination and J-1 difficulty parameters to characterize the extremity of the J ordered categories.

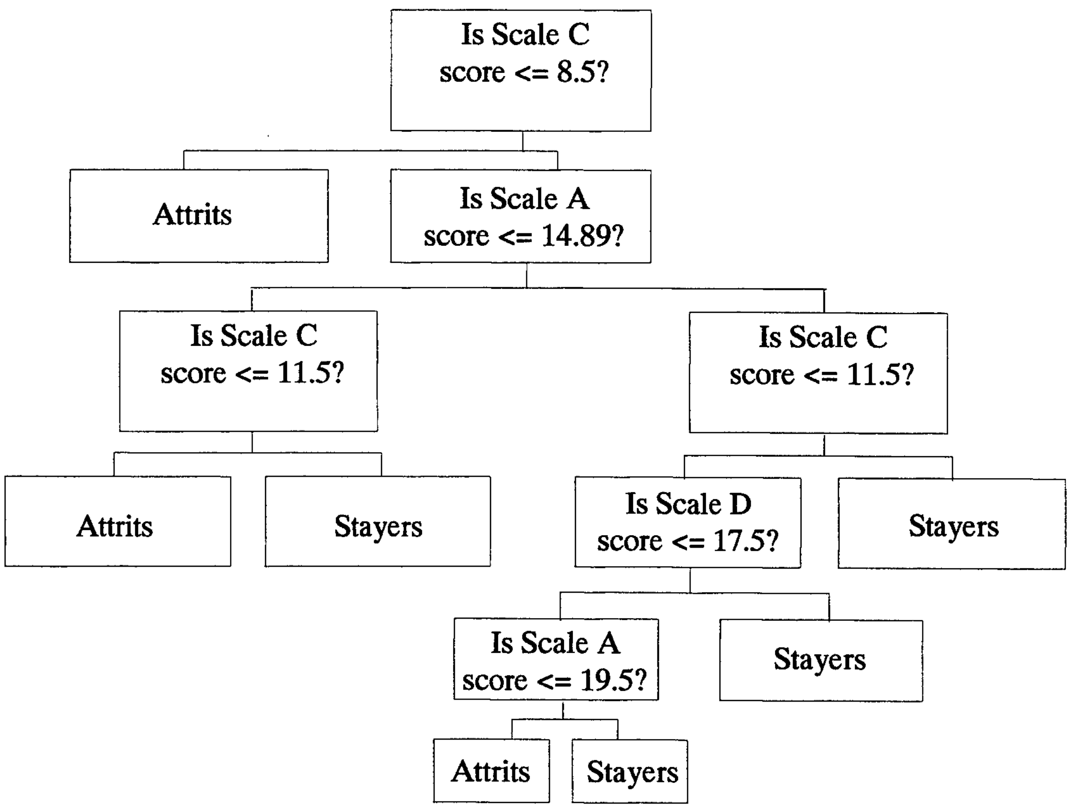


Figure 7.3. Classification tree with 7 terminal nodes.

Item response theory is difficult to describe in a few words. Readers are referred to Drasgow and Hulin (1990) for a relatively nontechnical introduction. Hambleton and Swaminathan (1985) and Langeheine and Rost (1988) provide more detailed descriptions.

Responses to the six AIM content scales were analyzed to determine if item response theory methods could improve prediction of attrition. Because these data were polytomously scored (with "options" 0, 1, and 2), Samejima's Graded Response (SGR) model was used to estimate item parameters separately for samples of nonattritees and attritees. After verifying the fit of the SGR model, using both graphical and statistical methods, the item parameters were used in optimal appropriateness measurement (OAM) analyses to classify respondents. The accuracy of the OAM classification procedure was examined using receiver operating characteristic (ROC) curves, which illustrate the proportion of hits at various false positive rates. A summary of the results of this investigation is provided below.

The application of item response theory for predicting attrition involves a three-step process: (a) calibration of the inventory, (b) examination of model-data fit, and (c) classification via OAM.

Calibration of the AIM Content Scales

Description of SGR Model

Because a single dominant dimension was found to underlie each of the six AIM content scales and the response data were scored polytomously (0, 1, 2), the SGR model was selected for item parameter estimation. For the SGR model, the probability of endorsing a response option, or category, depends on the discriminating power of the item (given by the a_i parameter) and the location of the difficulty parameters ($b_{i,j}$ and $b_{i,j+1}$) for that option on the latent trait (theta) continuum. The mathematical form of the SGR model is

$$\mathrm{P}(v_i = j|\theta = t) = \frac{1}{1+\exp[-1.7\mathrm{a}_i(t-b_{i,j})]} - \frac{1}{1+\exp[-1.7\mathrm{a}_i(t-b_{i,j+1})]},$$

where v_i denotes a person's response to the i^{th} polytomously scored item; j is the particular option selected by the respondent ($j = 1, \ldots, J$, where J refers to the number of options for item i); a_i is the item discrimination parameter and is assumed to be the same for each option within a particular item; b is the difficulty (i.e., extremity) parameter that varies from option to option given the constraints $b_{j-1} < b_j < b_{j+1}$, and b_J is taken as $+\infty$.

For stems having three options, as in the AIM scales, three parameters are estimated for each stem: one discrimination parameter that reflects the steepness of the option response function (ORF) and two difficulty (extremity) parameters that reflect the positions of the ORFs along the horizontal axis. As an example, the three ORFs for Item 1 of Scale B are presented in Figure 7.4. The discrimination parameter reflects the steepness of the ORFs; for example, a smaller parameter value would be reflected in flatter (i.e., less discriminating) ORFs. The difficulty parameters reflect where the ORFs are situated along the horizontal axis. For example, a smaller value of b_1 would move ORF1-0 to the left in Figure 7.4.

In Figure 7.4, the horizontal axis represents the latent trait (in this case, that measured by Scale B). The vertical axis represents the probability of endorsing a particular response option. The ORFs were computed using the equation for the SGR model shown above, where the discrimination parameter $a = 0.5$, and the two item difficulty (extremity) parameters $b_1 = -2.3$ and $b_2 = 0.9$. Notice that the ORF for the low (0) option is monotonically decreasing, whereas the ORF for the high (2) option is increasing. Also, at each value of theta, the probability values sum to 1.

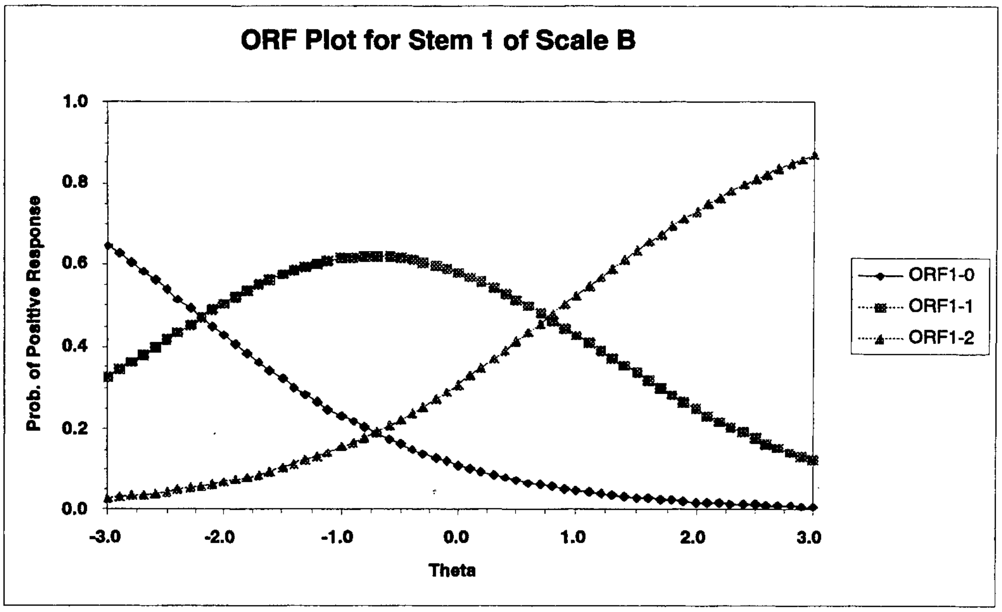


Figure 7.4. Representative option response function plot for SGR model.

Stem Parameter Estimation

Item parameters for the SGR model were estimated separately for the *total* samples of nonattritees (n = 18,016) and attritees (n = 4,521) using the MULTILOG computer program (Thissen, 1991). To facilitate convergence of the parameter estimation procedure, stems having negative or low (below .1) stem-total correlations were eliminated from the IRT analyses for both samples. The MULTILOG program was unable to estimate parameters for a few additional stems; consequently these stems were omitted from the analysis.

Examining Model-Data Fit

Graphical and statistical methods were used to examine the fit of the SGR model to the stems on each AIM scale for both nonattritees and attritees. This required that the total samples be split into calibration and validation subsamples; for each data set available for analysis, the odd-numbered cases were used for calibration and the even-numbered cases were used for validation. Stem parameters were reestimated for the calibration subsamples using MULTILOG. The validation subsamples were used for computing (a) empirical response functions (i.e., actual proportions endorsing each option) by the method described by Drasgow, Levine, Tsien, Williams, and Mead (1995) and (b) chi-square fit statistics.

Fit plots and chi-square statistics were computed using the MODFIT computer program (Stark, 2001). (See Drasgow et al., 1995 for a detailed description of the methods.) One fit plot was produced for each response option. As an illustration, Figure 7.5 presents results for the second stem in Scale C. The curves labeled ORF represent the theoretical ORFs for the second stem in Scale C computed using the parameters estimated from the calibration subsample for nonattritees. The points on the curves labeled EMP represent the empirical proportions endorsing each option computed using the cross-validation sample. The vertical bars in each plot represent the approximate 95% confidence intervals for the empirical proportions. In each plot, there is clearly a close correspondence between the ORFs and EMPs, which suggests that the SGR model fits the data well. Similar results were obtained for other AIM content scale stems. The fit plots were computed for all of the stems of the six scales. Note that the ORF and EMP curves are difficult to distinguish in Figure 7.5; this is good because it demonstrates excellent fit of the SGM to the AIM data.

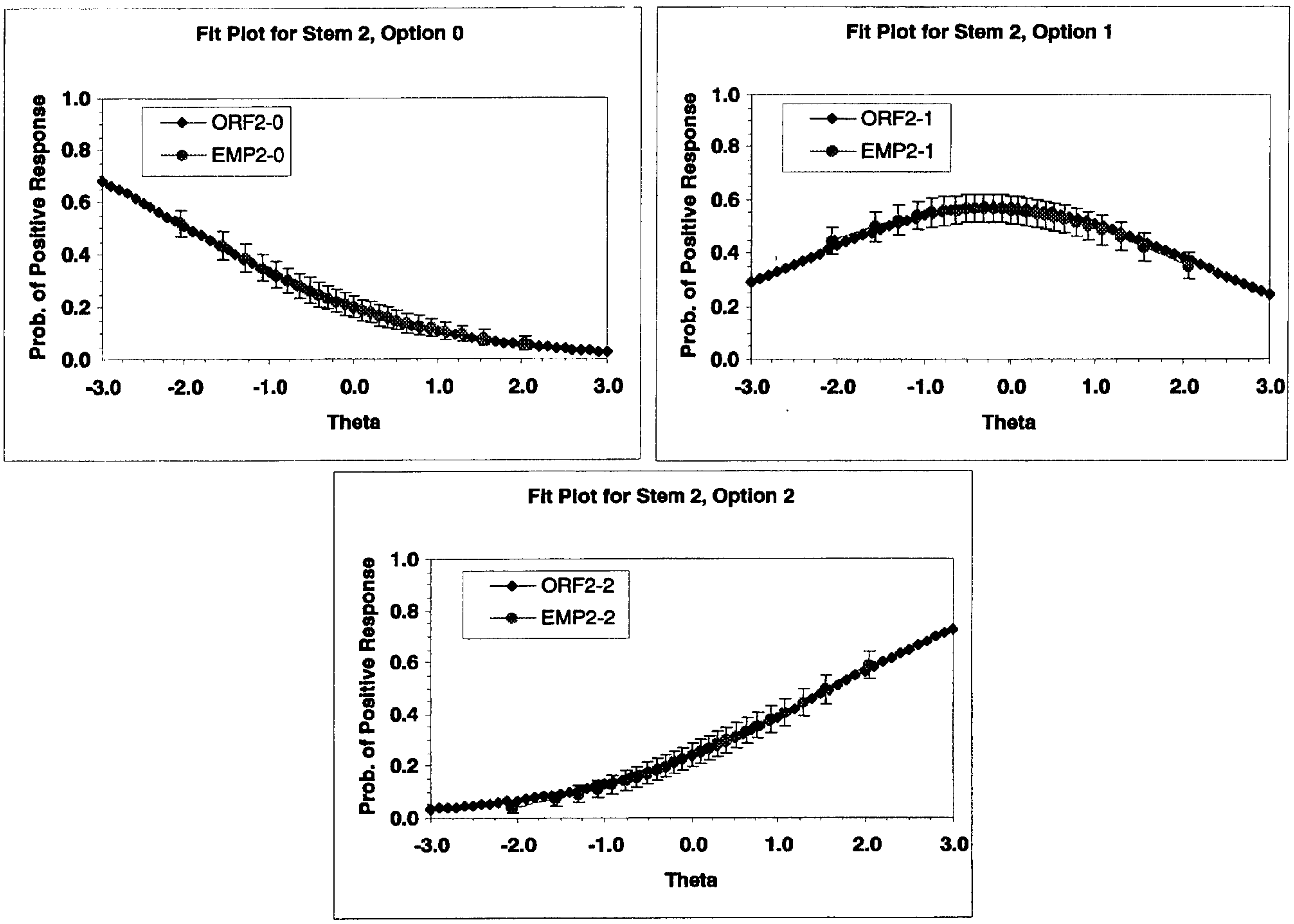


Figure 7.5. Example fit plots for Stem 2 in Scale C.

The investigation also examined model-data fit using chi-square statistics computed for individual stems, pairs, and triples of stems. Typically, this statistic is represented as

$$\chi_i^2 = \sum_{k=1}^{s} \frac{[O_i(k) - E_i(k)]^2}{E_i(k)},$$

where the χ^2 for stem i is computed from the expected and observed frequencies. The number of keyed options is given by s. $O_i(k)$ is the observed frequency associated with option k and $E_i(k)$ is the expected frequency of endorsing option k for a given IRT model. $E_i(k)$ is provided by the following:

$$E_i(k) = N\int P(v_i = k|\theta = t)f(t)dt.$$

Here, *f(t)* refers to the density of theta, which is assumed to be the standard normal, as ORFs are scaled with respect to that distribution. Because χ^2 fit statistics for single stems may be insensitive to violations of unidimensionality and to particular types of model-data misfit, χ^2 fit statistics should also be computed for combinations of pairs and triples of stems. These stronger tests for model-data fit are conducted by comparing the expected and observed frequencies in two-way and three-way tables in an extension of the above (see Drasgow et al., 1995).

Chi-square doublets and triplets were computed to detect violations of local independence and forms of misfit that are often missed by item singles. The chi-squares were also adjusted for a sample size of 3,000 and divided by their degrees of freedom to facilitate comparisons across samples of different sizes. According to Drasgow et al. (1995), adjusted ratios of chi-square to degrees of freedom of 3 or lower indicate good model-data fit. Table 7.4 presents the frequency distributions, means, and standard deviations of the adjusted χ^2/df ratios for the six AIM content scales.

The results in Table 7.4 indicate that relatively small χ^2/df statistics for item singles, doubles, and triples were obtained for all the AIM content scales. The average adjusted χ^2/df for single items ranged from 0.6 to 2.2; the average for doublets ranged from 2.4 to 3.7; for triplets the range was from 2.4 to 3.3. These results, in conjunction with the fit plots, indicate that the SGR model fit the AIM data well and could be used for classification of respondents based on OAM methods.

Classification Via Optimal Appropriateness Measurement

Optimal appropriateness measurement (OAM) provides the statistically most powerful methods for classifying examinees into two groups, such as nonattritees and attritees. These methods use a likelihood ratio test to classify examinees based on response probabilities computed under different psychometric models. Given appropriate models for the two types of responses, the Neyman-Pearson lemma states that no other method can be used on the same data to provide more accurate classification. Thus, the procedures are said to be optimal (Levine & Drasgow, 1988).

OAM can be used to classify a respondent into one of two groups based on the value of his/her likelihood ratio statistic. The likelihood ratio statistic is computed by dividing the marginal likelihood for attrition by the marginal likelihood of nonattrition. In this situation, we assume that the same process underlies response patterns for stayers and leavers, so the same marginal likelihood equation can be used for both groups. The only difference lies in the estimated item parameters used in the marginal likelihood equation shown below

$$Prob(v^*) = \int \left\{ \prod_{i=1}^{n} \sum_{j=1}^{J} \delta_j(v_i^*)P(v_i = j \mid t) \right\} f(t)\, dt\,. \qquad (1)$$

In the equation above, n is the number of stems in an AIM scale, t is an individual's standing on the latent trait, J is the number of response options for a stem i; $\delta_j(v_i^*) = 1$ if a score of v_i* was obtained in stem *i*, and 0 otherwise; $P(v_i = j|t)$ is the probability of a score of v_i on stem *i* given *t* (computed using the parameters for either nonattrition or attrition); and $f(t)$ is the normal density.

Table 7.4. Adjusted Chi-Square to Degrees of Freedom Ratios for Six Aim Content Scales

			Frequencies of adjusted (n=3,000) χ^2/df ratios								
Scale			<1	1<2	2<3	3<4	4<5	5<7	>7	Mean	*SD*
A	Nonattritees	Singlets	8	5	4	1	0	0	0	1.427	0.75
		Doublets	1	51	52	27	11	4	7	2.853	1.865
		Triplets	0	281	356	102	39	29	9	2.573	1.142
	Attrittees	Singlets	10	5	2	1	0	0	0	1.245	1.055
		Doublets	16	45	32	22	14	12	12	3.169	2.497
		Triplets	16	240	271	164	55	56	14	2.832	1.44
B	Nonattritees	Singlets	8	9	1	0	0	0	0	1.232	0.423
		Doublets	0	39	62	25	16	7	4	2.993	1.942
		Triplets	0	155	420	173	47	4	17	2.772	1.101
	Attritees	Singlets	7	5	2	0	0	3	1	2.236	2.719
		Doublets	13	31	29	25	18	22	15	3.738	2.78
		Triplets	5	145	238	235	116	58	19	3.256	1.494
C	Nonattritees	Singlets	1	6	0	1	0	0	0	1.65	0.677
		Doublets	0	8	10	6	3	1	0	2.751	1.112
		Triplets	0	16	31	8	1	0	0	2.433	0.569
	Attritees	Singlets	6	1	1	0	0	0	0	0.602	0.81
		Doublets	4	8	3	5	4	4	0	2.926	1.715
		Triplets	3	8	22	17	6	0	0	2.752	0.975
D	Nonattritees	Singlets	4	9	0	0	0	0	0	1.186	0.276
		Doublets	2	27	30	9	4	1	5	2.894	2.315
		Triplets	0	91	128	27	12	23	5	2.737	1.295
	Attritees	Singlets	8	3	0	1	0	1	0	1.152	1.697
		Doublets	14	19	16	12	8	4	5	2.845	2.245
		Triplets	17	78	101	53	18	16	3	2.702	1.37
E	Nonattritees	Singlets	5	6	3	1	0	0	0	1.564	0.761
		Doublets	1	39	20	18	8	10	9	3.478	3.293
		Triplets	0	115	159	96	42	27	16	3.18	2.007
	Attritees	Singlets	8	4	0	1	2	0	0	1.487	1.588
		Doublets	16	28	19	13	15	6	8	3.202	3.02
		Triplets	19	134	125	91	36	32	18	2.97	1.787
F	Nonattritees	Singlets	6	7	4	0	0	0	0	1.468	0.735
		Doublets	2	57	49	17	4	4	3	2.665	2.686
		Triplets	0	310	281	49	9	1	30	2.434	1.441
	Attritees	Singlets	8	3	4	0	0	1	1	2.239	3.441
		Doublets	26	33	29	16	11	16	5	2.907	2.864
		Triplets	33	257	220	112	25	10	23	2.533	1.531

As an example of the OAM procedure, consider the following. For responses to, say, Scale C, first, compute the marginal probability of a respondent's Scale C responses using the SGR item parameters for attritees. Second, compute the probability of the responses using the parameters for nonattritees. Third, compute the ratio of these two probabilities. Finally, if the ratio is large (i.e., the responses are better described by the model for attritees), predict that the respondent will attrit; otherwise, predict that the respondent will stay in the Army.

For each respondent we computed six likelihood ratio (LR) statistics (one per AIM content scale) using Stark's OAM computer program (Stark, 2001). As noted in the preceding paragraph, we computed LR values by the following process. Prob(v^*) given in Equation 1 was computed twice, once using item parameters estimated from the attrite sample and once computed using item parameters estimated from the nonattrite sample. The ratio was computed by placing Prob(v^*) from the attrite sample in the numerator and Prob(v^*) computed from the nonattrite sample in the denominator. Once all the likelihood ratios were obtained, we used logistic regression to determine the best linearly weighted sum of LR values for predicting the dichotomous nonattrition/attrition outcome. We then generated ROC curves for each AIM content scale and the logistic regression composite to examine how well the OAM procedure differentiated between groups of nonattritees and attritees. Figure 7.6 presents an example of a ROC curve for the Scale C. It can be seen that the OAM procedure differentiated nonattritees and attritees to a moderate degree. For example, for this scale, at a 20% false positive rate, about 33% of leavers were correctly identified. Similar results were obtained for the other content scales and the logistic regression composite. Those ROC curves are presented in Appendix C.

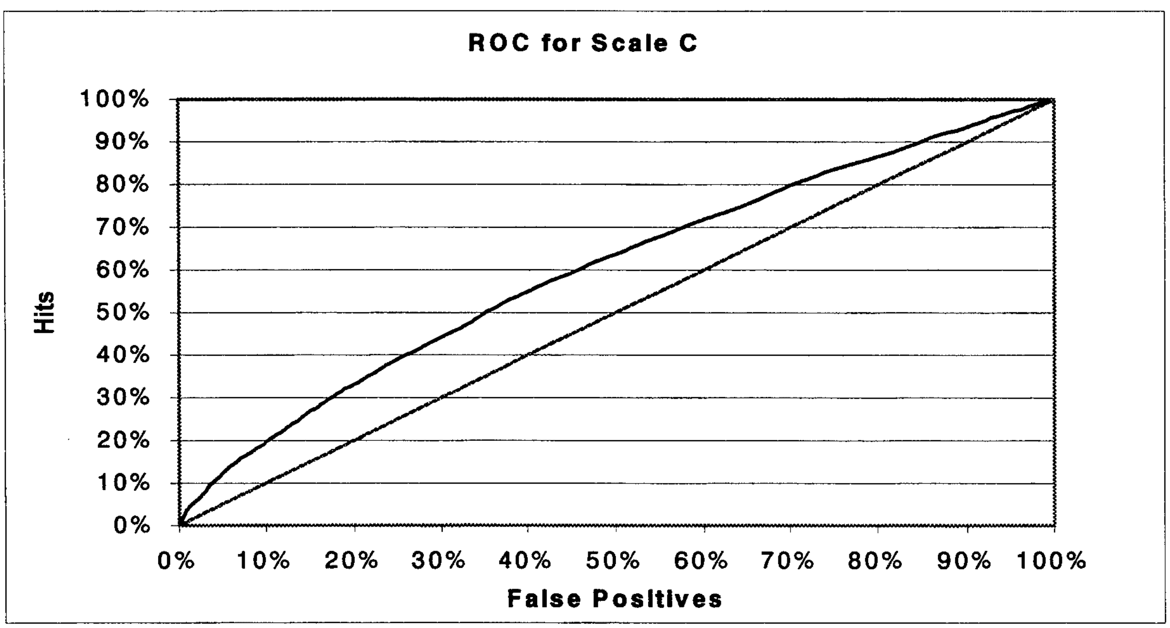


Figure 7.6. ROC curve based on likelihood ratio (OAM) values for Scale C.

Table 7.5 provides a summary of the OAM result at false positive rates of 10%, 20%, 30%, 40%, and 50%. The results in Table 7.5 indicate that the logistic regression composite provided the highest hit rates among the seven decision variables (six OAM-scored scales and the logistic regression composite). It correctly identified 22% of nonattritees at a 10% false positive rate, 35% of nonattritees at a 20% false positive rate, 45% at 30%, 56% at 40% and 66% at 50%. However, the logistic regression composite provided little improvement over Scales C and A alone. These results suggest that these two scale scores were most useful for predicting attrition using OAM. The worst classification rates were found for Scale E, which showed hit rates only about 5–10 percentage points higher than the corresponding false positive rates. For purposes of comparison, Table 7.5 also shows percent correct classifications for the current Adaptability Composite and the logistic regression equation.

Table 7.5. Percent of Correctly Identified Nonattritees Based on OAM Values for Six AIM Content Scales and the Logistic Regression Composite

		Percent false positive				
Scale		10%	20%	30%	40%	50%
A	Percent correct	22	33	44	53	64
B		18	29	39	50	58
C		20	33	45	55	64
D		17	30	40	50	60
E		15	27	37	49	59
F		17	29	40	50	59
Adaptability Composite	(Current Scoring)	17	27	41	50	59
Logistic Regression		19	29	44	52	60
Logistic Regression Composite		22	35	45	56	66

Summary and Discussion

This chapter reports a variety of analyses that provide insights into the predictability of attrition from the AIM scales. Analysis began by computing correlations between individual options (i.e., stems) of items and retention at 3, 6, 9, 12, 15, 18, and 21 months. Table 7.1 presents results of this analysis for Scale C. At 3 months, correlations of the individual options range from .053 to .082 with retention, which indicates that all of the options are positively associated with retention. The correlations tend to grow slowly as the retention variable matures, with option-retention correlations ranging from .090 to .146 at 21 months.

Stem-retention correlations do not vary widely within a particular AIM scale. This is demonstrated by the ranges of correlations cited in the preceding paragraph for Scale C. The other AIM scales show similar patterns.

Table 7.2 presents scale-retention correlations at 3, 6, 9, 12, 15, 18, and 21 months. Interestingly, Scale C appears to be slightly more correlated with retention than the Adaptability Composite. The correlations of scores on Scale A with retention are relatively stable across time (.114 at 3 months vs. .139 at 21 months) whereas the correlations of Scale B grow from .065 to .142 across time. Scores from Scales F and D have modest positive correlations at all time intervals, and Scales E and G are uncorrelated with retention.

Linear and logistic regressions were conducted to investigate whether optimal weighting of AIM scales could better predict retention than the Adaptability Composite. At 12 months, the Adaptability Composite correlates .132 with retention whereas the multiple correlation was .158 for the linear regression of retention on AIM scales. Thus, a modest improvement in prediction can be obtained by optimal weighting. Interestingly, the coefficient for Scale B was not significant and the coefficient for Scale E was significant but negative in sign. Scales C and A were the most important variables in the multiple regression equation. A similar pattern of results was obtained from logistic regression.

In the final portion of the analysis, two relatively exotic analyses were conducted: CART and OAM. The ROC curves summarized in Table 7.5 and Figure 7.7 show that CART and OAM provide modest improvements over the current Adaptability Composite in predicting retention. For instance, at a 20% false positive rate, the Adaptability score yields a 27% correct identification rate and the OAM composite yields a 35% correct identification rate. At a 30% false positive rate, the Adaptability Composite score has a 41% correct identification rate, the logistic regression equation has a 44% correct identification rate, and the OAM composite provides a 45% correct identification rate.

Another way to characterize the effectiveness of the alternative classification methods can be found in Table 7.6, which provides effect sizes and the percentages of attrits that "fail" and "pass" at the 10^{th} and 30^{th} percentiles. Note that the effect size for the Adaptability Composite is .342, the effect sizes for linear and logistic regressions are .405 and .433, and the OAM composite's effect size is .457. Again, these results show modest improvements can be obtained from alternative ways of predicting retention.

Table 7.6. Table of Effect Sizes and Implications of Cutoff Scores

Method (Scale)	Hits (Attrits below 10^{th} percentile)	False Acceptance (Attrits above 10^{th} percentile)	Hits (Attrits below 30^{th} percentile)	False Acceptance (Attrits above 30^{th} percentile)	Effect size
Adaptability	34.2%	18.2%	26.3%	16.6%	0.342
A	33.0%	18.1%	26.2%	17.0%	0.291
B	26.7%	18.8%	23.0%	18.2%	0.189
C	32.0%	18.0%	27.2%	16.5%	0.362
D	28.6%	18.7%	24.1%	17.9%	0.196
E	21.6%	19.5%	20.8%	19.3%	0.001
F	30.2%	18.6%	24.0%	17.9%	0.201
Linear Regression	37.4%	17.8%	28.2%	16.1%	0.405
Logistic Regression	37.2%	17.8%	28.1%	16.2%	0.433
A (OAM)	36.9%	18.2%	27.2%	17.0%	0.235
B (OAM)	30.9%	18.9%	25.3%	17.8%	0.257
C (OAM)	34.6%	18.5%	27.5%	16.9%	0.364
D (OAM)	29.4%	19.0%	25.2%	17.9%	0.185
E (OAM)	25.8%	19.4%	24.1%	18.4%	0.210
F (OAM)	31.1%	18.9%	25.5%	17.8%	0.266
OAM Composite	37.4%	18.3%	28.5%	16.5%	0.457

In sum, the analyses reported in this chapter indicate that the AIM is a well constructed instrument; only two stems were deleted due to unsatisfactory measurement properties. More complex alternative weighting schemes yield modest improvements in predicting retention.

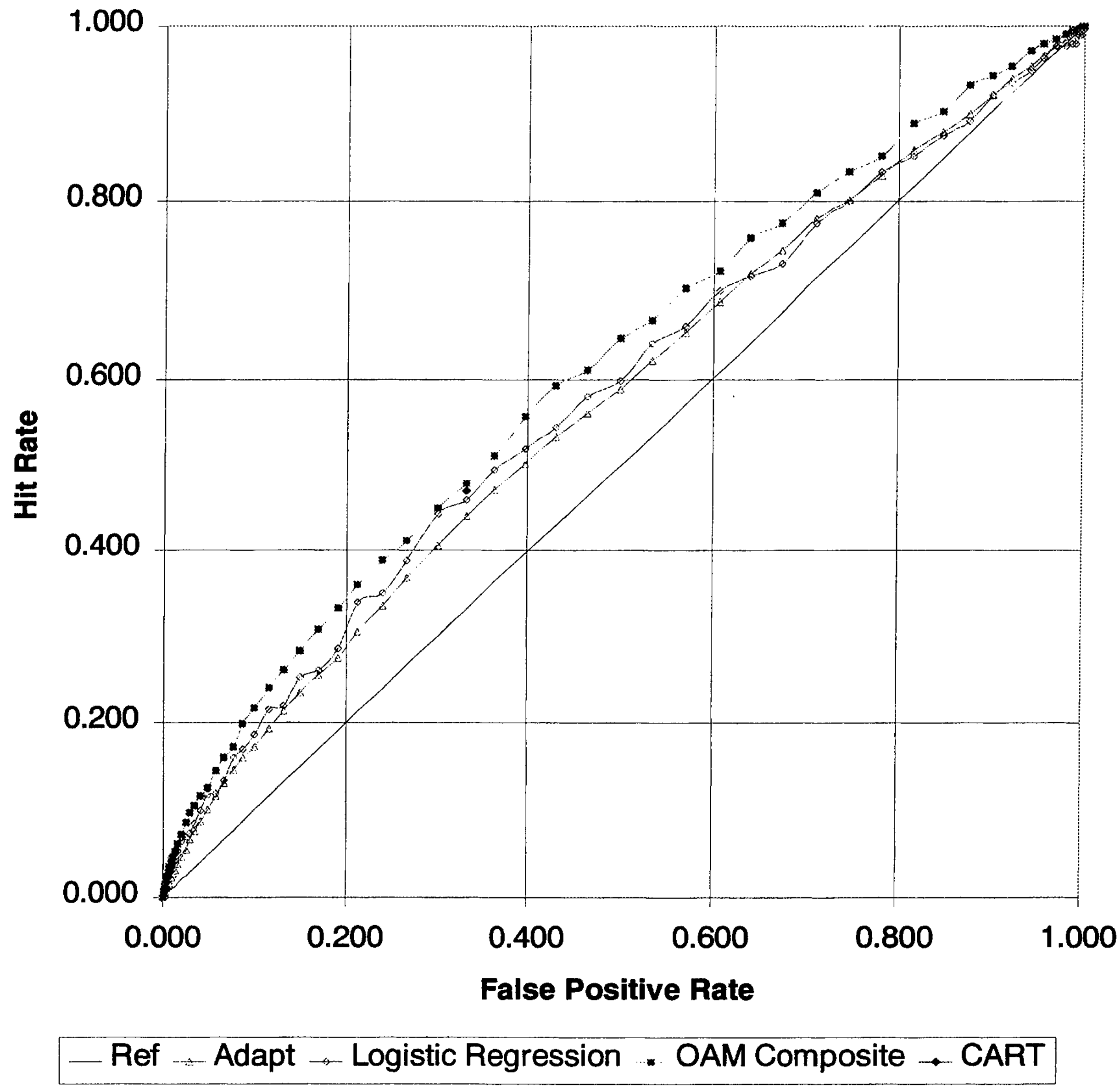


Figure 7.7. Comparison ROC curves.

CHAPTER 8. ROBUST MODELING AND OPTIMAL CLASSIFICATION FOR AIM

Michael V. Levine and Bruce A. Williams
University of Illinois, Champaign-Urbana

The following chapter describes the initial effort to explore the use of Multilinear Formula Scoring (MFS) for modeling item responses against attrition in the AIM research sample. It was later determined that the MFS model developed here did not fit the data from the operational sample. For this reason, a new MFS model was eventually developed using the operational data from GED Plus. Results from that work showed that MFS scoring significantly improved (approximately doubling) the validity of AIM against attrition in the operational sample, relative to the traditional scoring of the Adaptability Composite. Importantly, the validity of MFS AIM in the operational sample approached the validity of the Adaptability Composite that was observed in the research setting.

We believe that AIM's validity may continue to improve through further refinements of MFS modeling. Such refinements include (a) the effective identification of fakers (e.g., appropriateness measurement) and (b) the adjustment of scores for faking (e.g., robustification). Work is continuing in these areas.

Introduction

The purpose of the work reviewed in this chapter was to develop an item response theory (IRT) model for AIM. The specific objectives of the work were (a) predict attrition from AIM item responses at least as well as other methods; (b) approximate an upper bound for predicting attrition from AIM item responses; and (c) develop a model suitable for ameliorating the effects of faking (i.e., making false, socially desirable responses to AIM items).

Adapting IRT to AIM

Personality inventories in general pose a special problem for IRT because each inventory item, in addition to measuring what it is designed to measure, is measuring a person's motivation and ability to fake. AIM's complexity and unique format pose special challenges for IRT. Three problems and our solutions are briefly sketched below. The sections that follow provide details.

Item Format

Problem: Conventional IRT models allow only one response per multiple choice item. AIM's partially ipsative format obliges the respondent to choose a *Most Like Me* and *Least Like Me* response from the four alternative options.

Solution: Model each respondent as choosing one pair of responses to an item instead of making two separate choices. Thus each AIM item becomes a conventional one-response, 12-option item. Conventional "polytomous" IRT analyses can then be used. (Polytomous IRT techniques were described with references and illustrated in Chapter 7's 3-option SGR analyses.)

Multidimensionality

Problem: Efficient IRT model-fitting software is available only for unidimensional models. But each AIM item (by design) is sensitive to more than one dimension of individual differences.

Solution: Approximate a multidimensional IRT model for AIM with a unidimensional model. Levine (2001) has shown that multidimensional IRT models can be approximated with special unidimensional IRT models called *multilinear formula score* (MFS) models. There is reliable software available for fitting MFS models to item response data. Furthermore, MFS models have been observed to produce superior fits of personality inventory data (Chernyshenko et al., 2001; Zickar, 1997).

Item Transparency

Problem: AIM's partially ipsative format is not perfectly resistant to faking. Recruits evidently can guess which faked responses will raise Adaptability Composite scores.

Solution: Score AIM answer sheets with odds-in-favor-of-attrition scores. Odds scores are non-linear functions of item responses; falsifying a particular item response can raise or lower the respondent's odds score according to how s/he responded to other items. With odds scores, items become less transparent.

Optimal Classification

Chapter 7 introduced the concept of optimal classification in its discussion of the Samejima Graded Response (SGR) model and OAM. The concept is reviewed here to prepare the ground for two observations that are developed in the next two sections: (a) the ideal of nearly optimal classification is most likely to be achieved with robust modeling, and (b) nearly optimal prediction of attrition is possible before our understanding of AIM has progressed to a point where AIM can be used to accurately measure individual differences.

In the context of our research, "optimal classification" has a specific and narrow meaning. We are concerned with two theoretical probability distributions defined on a huge but finite sample space. The sample space has 12^{27} (= 1.3×10^{29}) points, one for each possible vector of item responses to the 27 AIM items.

One of the two distributions gives the relative frequency of item-response vectors for *nonattritees* (defined here as persons who remain in military service for at least 12 months). The probability assigned to a vector of responses by the nonattritee distribution is the probability of sampling a recruit having the specified vector of responses when randomly selecting a recruit from an extremely large pool of nonattritees. The second of the two probability distributions applies to *attritees* or recruits who withdraw early from military service. The probability assigned to a vector of item responses by the attritee distribution is the probability of randomly sampling the data of an attritee.

In sum, we have two number-valued functions defined on the set of all possible answer sheets, p_{attrit} and $p_{nonattrit}$. For any particular answer sheet, both p_{attrit} and $p_{nonattrit}$ will most likely be tiny numbers. However, it may turn out that for one person's answer sheet p_{attrit} is 10 or more times larger $p_{nonattrit}$. Commonsense (and mathematical statistics) say that an applicant producing an answer sheet that is much more probable among attrits is likely to turn out to be one of the recruits who attrit.

An *odds-based* decision rule is one of many possible ways to use a person's item responses to predict whether the person will leave the military or stay. When p_{attrit} and $p_{nonattrit}$ are given rather than estimated, an odds-based decision rule is obtained by (a) computing p_{attrit}, the probability of sampling the person's item responses according to the attritee probability distribution and (b) dividing it by $p_{nonattrit}$, the probability of sampling the person's item responses according to the nonattritee probability distribution, as follows:

$$\frac{p_{attrit}}{p_{nonattrit}}$$

An odds-based decision rule directs the decision-maker to predict attrition if and only the ratio is large. The ratio of the two probabilities is called here the "odds in favor of attrition" or simply the odds.

Odds-based decision rules are unusual among the ways psychologists use individual differences to predict behavior. Psychologists ordinarily use data to measure individual differences and then combine their measures to predict behavior. Odds-based decision rules are based on odds or the probability of observing item responses rather than measures of individual differences.

Chapter 7's SGR analyses combine features of both approaches. Measurement models are first applied to individual differences in the six AIM content areas to obtain an odds score for each of six ways to classify a person's response to an item. The odds scores are then combined by the methods ordinarily used to combine individual difference measures (regression methods) to obtain a number for predicting attrition.

Ordinarily it is necessary to estimate p_{attrit} and $p_{nonattrit}$ .When p_{attrit} and $p_{nonattrit}$ are estimated by fitting an IRT model to item response data, an odds ratio is approximated using the estimated p_{attrit} and $p_{nonattrit}$ distributions. In practical applications we are always dealing with an approximated odds ratio so goodness of fit tests are used to evaluate the approximation.

Odds-based decision rules have been known for some time (Neyman & Pearson, 1933) to be "unbeatable" or optimal. The standard "yes-no experiment" of signal detection (Green & Swets, 1966) is a typical setting in which odds-based predictions are demonstrably optimal. In the yes-no experiment, a person is sampled (with replacement) from one of two pools: an "event" pool and a "non-event" pool. A decision-maker uses information about the person to guess the pool from which the person was drawn. To evaluate the decision-maker, the experiment is repeated many times. Odds-based decision rules are optimal in the sense that a decision-maker using an odds-based decision will eventually tie or beat every other decision-maker.

With regard to the AIM, the two pools of interest are the attritee pool and the nonattritee pool, and the task is to use the item responses of a sampled person to decide whether the item responses came from a nonattritee or an attritee. The asymptotic proportion of hits (correct predictions of attrition) and proportion of false positives (incorrect predictions of attrition) are recorded. Odds-based decision rules are optimal in the following sense: For each possible false alarm rate, there will be a decision rule based on odds that will have the highest possible hit rate. A prediction scheme that is not odds-based can classify *as well as*, but not better than, odds-based rules.

In this chapter, "optimal classification" refers to classification with decision rules that are either odds-based or as powerful as odds-based decision rules. Truly optimal classification is possible only when the two probability distributions are known exactly. In practice, the two probability distributions p_{attrit} and $p_{nonattrit}$ are not known. They must be estimated from data.

Optimality and Robustness

To approach the ideal of optimal classification, it is necessary to estimate the p_{attrit} and $p_{nonattrit}$ distributions. IRT estimates p_{attrit} and $p_{nonattrit}$ by estimating the parameters of IRT models from sampled data. An odds-based decision rule can be close to optimal only if p_{attrit} and $p_{nonattrit}$ are well estimated.

Estimates of the p_{attrit} and $p_{nonattrit}$ distribution can be inaccurate for two reasons. The fist is sampling error. The models being fit to data may be sufficiently general in the sense that for some values of the model's fitted parameters estimated p_{attrit} and $p_{nonattrit}$ equal or are very close to the true p_{attrit} and $p_{nonattrit}$ distributions, but there are not enough data to accurately estimate the model's parameters. The second reason is bias or inconsistency. The model may make incorrect simplifying assumptions about p_{attrit} and $p_{nonattrit}$ and consequently, for all values of the model's parameters, estimated p_{attrit} and $p_{nonattrit}$ could be substantially different from true p_{attrit} and $p_{nonattrit}$.

IRT models can be roughly ordered from weak to strong. Weak models can fit complicated data sets but require large parameter estimation samples. Strong models can fit a narrower range of data sets, but their parameters can be estimated with small samples.

Robust modeling is modeling with weak models. In a mature area of psychometrics generally there are some strong IRT models that are known to fit typical data sets well. Typically, a robust model is formulated so that for special values of the robust model's parameters, one obtains the strong models that are conventionally used. MFS IRT models are robust in the sense that they can fit a wide variety of data sets and generalize the commonly used strong models. For example, the SGR models fitted in Chapter 7 can be approximated with any degree of accuracy by MFS models.

MFS models have been developed for fitting large, complicated data sets. The robustness of MFS models is superfluous for most data sets. As a rule, strong conventional models will fit samples of fewer than 3,000 people as well as or better than MFS. In simple, one-dimensional simulations of multidimensional models with the moderate to severe departures from the usual S-shaped item response functions, MFS typically requires samples of 6,000 to overtake conventional models.

MFS models have an unusual ability that make them especially valuable here. Although unidimensional, MFS models are able to fit item response data that have been generated by simulating many multidimensional IRT models. (For a discussion and formal proof of MFS's ability to fit some multidimensional data sets, see Levine, 2001.)

The robustness of MFS against departures from unidimensionality is of paramount importance in using AIM item responses to predict attrition. Each AIM item taps more than one content area. Additionally, individual differences in motivation, style, and inclination to dissemble—especially in operational data—are expected to magnify the need for robustness against multidimensionality.

Prediction without Measurement

A close examination of the logic of optimal classification leads to a surprising conclusion: Good measurement is not needed for good prediction. We observed that it was sufficient to have good estimates of p_{attrit} and $p_{nonattrit}$.

IRT modeling ordinarily focuses on measuring latent traits. This chapter describes a successful approach to estimating p_{attrit} and $p_{nonattrit}$ without measuring latent traits. MFS and some other IRT models lead to estimates of probability distributions and odds ratios that do not involve measuring latent traits.

This chapter is limited to predicting attrition rather than measuring individual differences and psychologically meaningful latent traits. The authors certainly agree that measuring psychologically meaningful latent traits is important. Regrettably, complicated calculations are needed to relate MFS's latent trait to psychologically meaningful individual differences. This is part of the price MFS pays for robustness against multidimensionality.

In collaboration with J. Douglas Carroll, the authors are actively engaged in applying multidimensional scaling to the problem of recovering latent traits from fitted MFS models. Eventually we expect to be able to use MFS for analyzing dimensions of individual differences and gaining insight into the causes of attrition.

Currently, MFS models item response data (including multidimensional personality inventory data) very well and measures meaningful latent traits badly. Neyman and Pearson have established that accurate modeling is sufficient for optimal prediction of attrition from item responses. Once again, this chapter is limited to prediction.

Using IRT to Estimate Answer Pattern Distributions

There are far too many possible answer sheets to directly estimate p_{attrit} and $p_{nonattrit}$ with relative frequencies. Most of the 1.3 x 10^{29} possible answer sheets are not observed in ARI's large sample. In fact it seems unlikely that any two recruits produced exactly the same answer sheet. Consequently, virtually all of the directly observed, sample relative frequencies are zero or 1/(the number of respondents).

IRT is used as a vehicle for estimating p_{attrit} and $p_{nonattrit}$ from the limited number of observed response patterns. As noted in Chapter 7, "Item response theory is difficult to describe in a few words." MFS, which builds upon the concepts and assumptions of basic IRT, requires even more words. Readers are urged to review Chapter 7's brief introduction to IRT, especially the implementation of the SGR model. MFS is introduced in this section by contrasting MFS and SGR. MFS and SGR use IRT in the same way to estimate asymptotic relative frequencies of item response patterns too numerous to directly estimate.

Option response functions (ORFs) are used to relate individual differences in a latent trait (theta) to item responses. Figure 7.4 displays SGR model ORFs for a typical item. The equation preceding Figure 7.4 gives the mathematical form of the ORFs. The number t or abscissa theta represents the latent trait (say, leadership). The precise shape of each of the ORFs is determined by the discrimination parameter a_i and the location parameters $b_{i,j}$. For SGR, the latent trait and each of the ORF parameters have a psychological interpretation.

ORFs allow one to estimate the asymptotic relative frequencies of answer sheets from samples much smaller than the number of unique possible answer sheets. The total number of parameters needed to specify a full set of leadership ORFs is far less than $3^{27} = 7.6 \times 10^{12}$, the number of answer patterns for 27 items. A standard statistical method (maximum likelihood estimation) is used to efficiently estimate the ORF parameters. Chapter 7's formula (1) can now be used to calculate estimates of the relative frequency of each of the possible leadership response answer patterns. Thus IRT is able to estimate relative frequency of answer patterns from a much smaller number of answer patterns than the number that are possible.

MFS's estimates of relative frequencies in broad outline follow the same pattern. ORFs are estimated by fitting item parameters. Chapter 7's formula (1) (page 7-11) is used to calculate estimated relative frequencies from estimated item parameters. Table 8.1 contrasts MFS and SGR.

Table 8.1. Comparison of MFS and SGR Approaches

Contrast	SGR	MFS
Shape of ORFs	monotone or unimodal	arbitrary and multimodal
Number of parameters	Small	Large
Parametric form	Non-linear	Linear
ability distribution	Normal	Arbitrary
Number of options/item	3	12
Content Areas	Modeled one-by-one	All modeled simultaneously

Both SGR and MFS are one-dimensional IRT models. The role of the one dimension is the fundamental difference between SGR and MFS. For SGR, theta is an unobserved latent trait to be measured and interpreted. On a superficial level, MFS's one dimension is a mathematical convenience introduced to simplify approximating p_{attrit} and $p_{nonattrit}$ and their ratio. On a deeper level, MFS's dimension is much more. Detailed information about the individual differences relevant to attrition is implicit in MFS's one-dimensional ORFs as surely as one-dimensional strands of DNA contain information about eye color, hair texture, and other physical characteristics (see Levine, 2001 for a detailed discussion of this point.)

MFS is fitted to item response data by a suite of computer programs collectively called ForScore. ForScore subroutines dynamically adjust the shapes of ORFs for each application. ForScore uses an initial set of ORFs to calculate a large set of similar ORFs. It uses a principal components analysis to represent the set of ORFs as linear combinations of a relatively small number of functions. By adjusting the coefficients of linear combination, ForScore selects ORFs that fit the data better than the initial ORFs. ForScore replaces the initial ORFs with the better fitting linear combinations. It then treats the new ORFs as initial ORFs and estimates a set of ORFs that fit the sampled data even better. The process is continued until the new ORFs give a negligible improvement of data fit.

Although MFS's ORFs have many parameters to estimate, the parameters enter the formulas for ORFs linearly. Linearity plays an important role in MFS and ForScore. Linear calculations are fast, numerically stable, and well-researched. Thus ForScore can fit very large, complicated data sets with an inexpensive desktop computer.

Linearity allows ForScore to control the shape of its fitted functions when it is desirable to do so. Many psychological intuitions can be expressed in the form of linear equations and linear inequalities about MFS ORF parameters. ForScore can use these equations and inequalities to find the best fitting monotonic, unimodal, S-shaped, bi-modal, or bell-shaped functions. In this research, linear inequalities are used to keep ForScore's ORFs between zero and one.

ForScore can use its estimated ORFs to find a best fitting distribution of theta (or, more precisely, a probability density function for the probability distribution of MFS's latent trait). A density can be fitted from the total sample of respondents or from any selected subsample of respondents. This feature (as explained in a later section) allowed us to estimate p_{attrit} almost as well as $p_{nonattrit}$ even though the attrition sample was relatively small.

ForScore estimates ORFs for each of the 12 options. ForScore's density-estimating program was used to estimate a probability density separately for the attritees and others. Formula (1) of Chapter 7 was then used to calculate an approximation of p_{attrit} and $p_{nonattrit}$ and their ratio, the odds score. Later sections provide some additional details about these approximations.

Estimating ORFs

MFS and ForScore represent ORFs as linear combinations of a basic set of functions. ForScore estimates ORFs by computing estimates (maximum likelihood estimates) of the coefficients of linear combination. An initial set of basic functions is derived from a strong IRT model or an educated guess about the shapes of ORFs. As ForScore improves its fit of data, ForScore revises its set of basic functions.

For the analyses reviewed in this chapter, ForScore was initialized by computing the Adaptability Composite score for each respondent. An intial set of ORFs was obtained as "smoothed regressions" on the Adaptability Composite score. In other words, for each AIM item, we calculated 12 histograms, one for each of the possible *Most Like Me, Least Like Me* response pairs. Each histogram gave the relative frequency of recruits choosing an option. A standard statistical technique (kernel smoothing) was used to convert the histograms into smooth functions of a continuous variable, theta.

To convert the histogram for a typical item option, say, item 1, response option <A is most like me, B is least like me> into a smooth function of a continuous variable, we replaced the histogram of item 1, option choice <A,B> by a continuous function as follows. For the number theta = t, the function value was the weighted average or integral of histogram values. The weights were proportional to the normal density with mean t and small variance so that histogram values close to t had large weights and histogram values far from t had small weights. Thus the initial ORFs were smooth (in fact, differentiable) functions taking values between zero and one with roughly the same shape as the Adaptability Composite score histograms.

ForScore generated a new set of ORFs containing the original smoothed histograms and many related functions. The enlarged set of ORFs includes all the ORFs that can be obtained as linear combinations of the smoothed histograms. The enlarged set of functions that ForScore will search to find better fitting ORFs also includes products and sums of products of the initial functions. From this enlarged set, ForScore selects ORFs that fit the sampled data better than the ORFs used to generate the enlarged set of functions. As noted in the previous section, ForScore iteratively replaces a current set of ORFs by better fitting ORFs until no further improvement of data fit is observed.

Generally 12 ORFs are estimated for each item. However, some response pairs were chosen by very few recruits. For each item, the rarely chosen response pairs, if any, were combined to define a "disjunctive" option which was scored as chosen if any one of the item's low frequency pairs was chosen. We also experimented with other groupings of item response pairs to form disjunctive options.

The same ORFs were used for all respondents, both attritees and non-attritees. We reasoned that the same individual differences were determining item responses in the same way for both groups of respondents.

By using a common set of ORFs for both groups, we avoided having to estimate ORF parameters from the relatively small sample of attritees. Additionally, we were able to analyze the data from recent recruits before enough time elapsed to determine whether the recruit would become an attritee. Of course, data from some of the recruits were set aside for later evaluation. At no time did ForScore have access to the hold-out evaluation sample.

Computing an Odds Score

To approximate p_{attrit} and $p_{nonattrit}$ and calculate an odds score, it is necessary to incorporate the ORFs in one model for attritees and a second model for non-attritees. Both models use the same ORFs and the same latent variable theta. The only difference between the two models is the probability distribution for theta.

We estimated a probability density for theta for the attrition model from a sample of attriting recruits. We estimated the probability density for the second model from a sample of recruits who did not withdraw early from military service. Some details on the estimation of the densities and their use to compute a score follow.

Recall that MFS and ForScore represent ORFs as linear combinations of a basic set of continuous functions. ForScore updates the basic set of functions as the computation proceeds and represents probability densities as linear combinations of the same functions used to represent ORFs. Densities are estimated by estimating the coefficients of linear combination. (Linear constraints on coefficients are used to guarantee that estimated densities are non-negative and that the estimated densities integrate to one.) ForScore's calculation thus yields a pair of estimated densities f_{attrit} and $f_{nonattrit}$ on the same theta scale as the estimated ORFs.

The estimated densities are substituted in formula (1) of Chapter 7 to obtain estimates of p_{attrit} and $p_{nonattrit}$. To be more explicit, the formula expresses the theoretical probability of sampling an answer sheet with the vector of item responses v^* as the integral of the product of two functions, $\int L\, f\, dt$. The first function $L = L(v^*, t)$ is the product of ORFs inside the curly brackets. The second function $f = f(t)$ is the probability density for a unidimensional IRT model. In our calculation of estimates of $p_{attrit}(v^*)$ and $p_{nonattrit}(v^*)$, the estimated ORFs are substituted in the formula for L and f is one of the two estimated densities f_{attrit} and $f_{nonattrit}$. The score used for an answer sheet is the ratio of these estimated probabilities.

Progress Towards the Three Goals

MFS odds scores have been predicting attrition from AIM item responses as well as or better than all other quantitative methods applied to AIM data. We anticipate this generalization being confirmed in projected future analyses of AIM data collected under operational conditions. We reason that the incidence of faking should be higher under operational conditions. In addition, we suspect that attritees and non-attritees reliably differ along a candor individual difference dimension. We conjecture that MFS's robustness against multidimensionality will give MFS odds scores an advantage when competing against strong unidimensional IRT models.

Our modeling of AIM data and estimation probabilities attempts a compromise between bias and sampling error. A suggestion originated by ARI indicated that sampling error currently separates us from optimality. Sampling error is decreased by estimating fewer parameters. Acting on ARI's suggestion we repeated the analysis leaving out some items that ARI suspected had little value in predicting attrition. Validity increased slightly.

Our first goal seems to have been achieved. MFS odds scores predict attrition from AIM item responses at least as well as other methods.

It may be possible to further improve predictability of attrition with odds scores by augmenting the MFS IRT model used to compute odds. Several likely indicators of attrition are ignored by current MFS odds scores. These include continuous objective variables such as age and weight/height ratios. When these variables are sorted into a small number of categories such as *young, average, old,* or *light, average, heavy* they become formally like multiple choice test items. We are currently attempting to improve prediction by incorporating them in an MFS IRT model and using them along with AIM responses to compute an odds score.

Our second goal was to approximate an upper bound for the predictability of attrition using AIM scores. As discussed above, it follows from the Neyman-Pearson lemma that odds scores are optimal when they are computed from p_{attrit} and $p_{nonattrit}$ distributions. MFS typically models large personality inventory data sets well (and better than strong unidimensional models).

Table 8.2 summarizes some goodness of fit results. Chapter 7's conventions for chi-square statistics were used. In this run, the 12 options were grouped to form disjunctive option-response categories. Options were grouped together if they involved exactly the same choices on the three most diagnostic content areas. Options were also grouped together if their smoothed option-score regression functions were very close together.

Table 8.2. Chi-Square/df per 3,000 Examinees

Value/Type	≤ 1	<1≤2	<2≤3	<3≤4	<4≤5	<5≤6	<6≤7	>7	Mean	*SD*
Single										
Nonattrits	10	13	1	1					1.20	.58
Attrits	13	3	3	2	3	1			1.46	1.90
Pairs										
Nonattrits	18	252	20	6	2	1	1		1.53	.645
Attrits	58	140	67	12	6	0	1	1	1.85	1.14
Triple										
Nonattrits	32	2,210	53	5					1.31	.26
Attrits	465	1,365	405	60	2				1.53	.65

MFS's close fit of AIM item responses strengthens our belief that MFS odds scores make nearly optimal use of AIM data. Our MFS studies indicate that when AIM item responses are used exclusively, predictions based on AIM item responses can be correct about 62% of the time in a two-alternative forced-choice experiment. As noted above, increases in predictability of attrition can be obtained by using AIM item responses along with other indicators of attrition.

Our final goal was to develop a model suitable for ameliorating the effects of faking. Projected analyses of data having a higher incidence of faking are needed to see whether the diminished transparency of odds scores is effectively ameliorating the effects of faking. In addition, as noted above, it seems likely that there are reliable individual differences in candor among attritees and non-attritees that can be incorporated in an MFS mode to further protect against faking.

Unidimensional IRT models have provided counters to testing anomalies in ability and achievement testing. MFS IRT models now fit personality inventory data about as well as IRT models have been fitting ability and achievement data. It is now advisable to attempt to adapt the methods developed for cognitive data to personality inventory data, at least when large samples of item response data are available. Although formally unidimensional, MFS and ForScore provide an accurate, item-by-item account of item responses determined by diverse individual differences.

CHAPTER 9. RECAP OF POST-IMPLEMENTATION INVESTIGATIONS

Deirdre J. Knapp, Rodney A. McCloy, William J. Strickland, and Brian K. Waters[1]
Human Resources Research Organization (HumRRO)

In this last chapter, HumRRO highlights the major findings of the efforts conducted from September 2000 through July 2001. However, important developments have been made in ARI's AIM program since that time. In the postscript that follows, we provide an update on these newer developments.

In this last chapter, we summarize the major post-implementation findings regarding the AIM. We then discuss technical questions associated with the partial ipsativity of the AIM and close with some thoughts about directions for future work. Consistent with the goals of the current project, this discussion focuses primarily on the use of AIM as an enlistment screening tool. Much, but not all, of the discussion also is relevant, however, to the use of AIM in other contexts (e.g., to predict performance in certain military occupational specialties or at higher organizational levels).

Summary of Post-Implementation Investigations

This report has documented work to investigate several areas related to the use of the AIM as an enlistment screening tool. After documenting the analysis datasets in Chapter 2, the remaining chapters described work related to the following areas:

- Validation results using operational data (Chapter 3)
- Scaling of alternate forms (Chapter 4)
- Investigation of supplemental predictors of attrition (Chapter 5)
- Fairness analyses (Chapter 6)
- Alternative scoring methodologies (Chapters 7 and 8)

Each of these investigations is briefly summarized in this section.

Validation Results

Chapter 3 provided a preliminary assessment of the validity of the AIM administered operationally (as part of the Army's GED Plus program) for predicting 3-, 6-, and 9-month attrition. The results showed a drop in validity from that previously observed in research samples. The AIM was somewhat effective in reducing 3-month attrition; the use of AIM reduced attrition by an estimated 9.7% over the attrition rate observed in the comparable research

[1] The authors wish to acknowledge the valuable assistance of Dr. Eric Heggestad (Colorado State University) in the preparation of this chapter.

sample (i.e., educational Tier 2 [non-high school diploma graduate] soldiers). AIM did not decrease the expected rate of attrition at 6 months or 9 months.

Scaling Alternate AIM Forms

Two pairs of alternate AIM forms (Initial Forms A and B and Revised Forms A* and B*) had been developed in prior AIM work (Heggestad, Young, Strickland, & Rumsey, 1999). Prior to scaling, we evaluated the relative effectiveness of the two sets of alternate forms with the understanding that the better set (either A and B or A* and B*) would be scaled to the original form. As described in Chapter 4, McCloy, George, and Reeve compared the two sets of alternate forms by examining rank-order correlations, decision consistency, score differences, and predictive validity. Given that neither set of forms was substantially better than the other based on these analyses, both sets of forms were scaled to the original AIM. Thus, Chapter 4 reports preliminary scaling of Forms A, B, A*, and B* on the research sample. Pending future decisions related to the scoring of AIM (e.g., whether the Adaptability Composite will be revised), the scaling exercise will need to be revisited.

Investigation of Supplemental Predictors

In Chapter 5, Putka and McCloy merged the Army AIM research data with data collected from the Army's First Term attrition research project. This allowed the investigation of variables that could be used in conjunction with AIM to increase our ability to predict attrition. Analyses were conducted separately for two groups at high risk for attrition – Educational Tier 2 recruits (both male and female) and female Tier 1 (high school diploma graduate) recruits.

The results suggest that the Army may have much to gain by further exploring the Soldier Reception Survey (SRS) variables identified here as potentially salient predictors of 18-month attrition status. While the combination of AIM and current operational variables was moderately predictive of attrition status, SRS variables increased both the validity and utility of the resulting models. Specifically, the expected decrease in attrition achieved by screening out 10% of Tier 2 soldiers based on a composite comprising AIM and operational variables was estimated to be about 4%, whereas adding SRS variables to the same composite yielded an expected drop of 11%.

Fairness Analyses

Sackett and Laczo at the University of Minnesota reported AIM fairness analyses (adverse impact, differential validity, and differential prediction) in Chapter 6. They used both research and operational data from the Army, as well as research data from the Air Force. Across all samples very small mean differences were observed between subgroups on the AIM scales or the selection composite. As such, the AIM should not result in adverse impact (defined by the four-fifths rule) against any of the subgroups when used to screen out low-scoring individuals. Although AIM scale and composite scores were much higher in the operational setting, subgroup differences remained essentially the same.

Sackett et al. raised an important methodological consideration when examining the differential validity of the AIM: whether the attrition criterion is examined in either a cumulative or non-cumulative fashion. In the cumulative analyses, the sample comprised only those soldiers who could potentially have stayed for the applicable time in service. The mean AIM score from soldiers who attrited by the applicable time period (e.g., 21 months) was compared to the mean AIM score from soldiers who stayed. In the noncumulative analyses, the mean AIM score of individuals still in the Army at a given time point (e.g., 21 months) was compared to the mean of individuals who left in the preceding 3 months (e.g., who left within 19-21 months). These alternative approaches (of which the first is most often used by researchers) ask different but important questions.

With regard to the differential validity analyses, there were no gender differences in any of the three datasets analyzed. The Army research and operational datasets yielded no systematic differential validity across race subgroups. The Air Force data, however, showed Black-White differences consistently across attrition time periods. Across all samples and subgroups, validity (as indexed by effect sizes on the AIM scales when comparing attritees and non-attritees) was highest for the initial 3 months of service, and declined after that. When examining data in a cumulative fashion, the decline was slight. When examining data in a non-cumulative fashion, however, the decline was quite dramatic: In the Army sample, validity (i.e., effect size) dropped from –.40 to –.20 to –.11 across the three time periods (3 months, 6-12 months, 15-24 months); in the Air Force sample, validity dropped from –.53 to –.11 to –.06 across the same three time periods despite increasing turnover baserates.

The AIM exhibited differential prediction against Blacks and Hispanics, and the use of a common regression line showed predictive bias against members of these groups. In particular, it was found that members of these subgroups were less likely to leave the service than AIM would predict. It is not clear whether the AIM is truly biased, however, or whether the obtained results are due to the omitted variables problem. This problem occurs when the effects of variables omitted from the regression analysis cause researchers to reach potentially misleading conclusions. In this case, variables correlated with race/ethnic group (e.g., socioeconomic status, perceived labor market alternatives, cultural values regarding completing a tour of duty) but not included in the analyses may explain the differential prediction that on the surface appears to be due to race. The data available for examination here did not include likely omitted variables. If key omitted variables were identified, and if their inclusion in the model eliminated the race effect, one would conclude that the AIM itself did not have predictive bias. But these omitted variables would have to be included along with the AIM in a selection system for that system to be free from such bias.

Given these findings, it is also important to note that the legal requirement to investigate differential validity only kicks in when a test shows adverse impact. As the AIM Adaptability Composite score shows very small subgroup differences, and hence no adverse impact, there is no legal obligation to conduct validity or differential prediction analyses.

Alternate Scoring Methodologies for AIM

Dragow and his colleagues at the University of Illinois investigated several alternate methodologies for predicting attrition from the AIM scores. Besides traditional test score and regression analyses, they explored the application of two recent techniques for analyzing AIM-type data: Classification and regression trees (CART) modeling of attrition with the AIM scales at 12 months of service, and classification via item response theory (IRT) optimal appropriateness measurement (OAM). These results are reported in Chapter 7. Chapter 8 documents research conducted by Levine and Williams on the application of a multidimensional robust modeling strategy for scoring the AIM.

Classification and Regressions Trees (CART)

CART identifies pass/fail points on a series of selected AIM scales to optimally classify individuals as attritees or nonattritees. A "decision tree" may have several if/then branches. For example, if one's Work Orientation score is greater than 15, then predict nonattrit and move to the next branch. Otherwise, predict attrit and the examinee fails the screen. For those who pass the first branch, a second branch is considered. For instance, the second branch may have a cut score requirement on the Adjustment scale; those who "fail" this second cut-score are removed from consideration and those that "pass" continue on to the next branch. These terminal nodes correspond to the categories of the outcome variables. The CART program exhaustively examines all possible binary splits with each predictor and arranges them into separate trees with the best predictor being the "root" of the trees. CART then chooses the best tree based on the costs (misclassification rates) associated with the outcome variable (12-month attrition in this case). Furthermore, CART selects trees so as to minimize the number of terminal nodes that consist of only a handful of cases. CART also cross-validates the "trees" against a holdout sample. The CART analysis yielded 39 trees ranging in complexity from 2 to 2,802 terminal nodes. Characteristics of five of the best trees compare favorably with the current Adaptability Composite.

OAM – AIM Prediction Using Item Response Theory

OAM provides the statistically most powerful methods for classifying examinees into two groups, such as nonattritees and attritees. These methods use a likelihood ratio test to classify examinees based on response probabilities computed under different psychometric models. No other approach can be used on the same data to provide more accurate classification. Thus, the procedures are said to be optimal.

Responses to the six AIM content scales were analyzed to determine if IRT methods could improve prediction of attrition. Because AIM items are polytomously scored, Samejima's Graded Response Model (SGR) was used to estimate item parameters separately for samples of nonattritees and attritees. After verifying the fit of the SGR model, the item parameters were used in OAM analyses to classify respondents. The accuracy of the OAM classification procedure using AIM data was examined using the proportion of hits at various false positive rates as the criterion (plotted as ROC curves).

The OAM procedure showed some improvement over the current Adaptability Composite. For example, at a 30% false positive rate, the Adaptability Composite had a 40% correct identification rate while the OAM composite had a 45% correct identification rate.

Robust Modeling and Optimal Classification for AIM

This item-level model for the AIM is being developed primarily to support countermeasures to false responding and improved prediction of attrition. Levine and Williams used weak multidimensional nonparametric models to analyze AIM data. To use standard IRT modeling methods, each AIM predictor item was analyzed as a 12-option, multiple-choice item. The 27 items were called "predictor items" since they were being used to predict attrition. The criterion (12-month attrition) was used as a 28th (dichotomous) item. The AIM was analyzed as an ordinary, multiple choice test. This was done by treating each pair of responses (*Most Like Me, Least Like Me*) as a single-option choice. In this way, the AIM's partially ipsative format could be simulated as a simple, polytomous multiple choice format. Predictions based on this item-level model were found to be at least as accurate as predictions made with any other procedure that has been used with the AIM data.

This work integrated the Samejima Graded Response Model (SGR) results from Chapter 7 and 12-option, simultaneous modeling of several scales. The two methods predicted attrition about equally well. The trichotomous SGR functions were more likely to provide useful feedback to item writers. The 12-option models seemed better suited to appropriateness measurement and robustification. Levine's work is promising.

Cross-Chapter Summary

To enable a more direct comparison of the AIM scoring methods discussed in this report, Table 9.1 shows hit rates (i.e., the proportion of attriting soldiers who would have been screened out by the Adaptability Composite) that were calculated for each approach at five false positive rates. The false positive rate is the proportion of soldiers who would have stayed had they not been screened out by the Adaptability Composite. Note that these results are based on data collected in a research setting. As demonstrated in Chapter 3, it is unlikely that hit rates would be as high in an operational sample.

Table 9.1. Hit Rates at Five False Positive Rates for Four Alternative Approaches to Scoring the AIM

False Positive Rate	Original Adaptability Composite	Logistic Regression Composite (Current Scoring)	OAM Logistic Regression Composite	Levine's Composite
10	17	19	22	22
20	27	29	35	36
30	41	44	45	46
40	50	52	56	57
50	59	60	66	66

As demonstrated by Receiver Operating Characteristic (ROC) curves, the hit rates vary proportionally with false positive rates (see Chapter 3, Figure 3.1). As illustrated in Table 9.1, optimally weighting the components of the current Adaptability Composite can improve its predictive power (compare Original Adaptability Composite with Logistic Regression Composite [Current Scoring]). Applying OAM procedures to the composite further increases its predictive power (from 5 to 9 percentage points). The OAM procedure and Levine's approach yielded nearly identical results.

Technical Issues Rooted in the Partial Ipsativity of the AIM

Analyses of AIM data from both research and operational settings have surfaced technical issues emanating from a common source: The partial ipsativity of AIM scores that results from its scoring strategy. In this section, we discuss these issues and suggest why and how the Army should consider addressing them. We should note that this discussion is relevant to any operational use of AIM, not just the GED Plus application.

Partial Ipsativity

Test scores can range along a continuum from *normative* (a respondent's score is statistically dependent on the scores of other respondents in the population of interest but statistically independent of the respondent's other scores) to *ipsative* (a respondent's score is statistically dependent on the respondent's other scores but statistically independent of—and not comparable with—the scores of others in the population). Hicks (1970) cited several characteristics of ipsative scores. Measures that embody some of the characteristics of ipsative tests (e.g., a higher score on a particular trait results in a lower score on other traits), but not others (e.g., there is some variability in respondents' total scores), are termed *partially ipsative.*

An important difference between normative and ipsative scores concerns the types of inferences one may draw from them. Specifically, a normative measure provides estimates of each respondent's true score on each trait that the measure assesses. Thus, respondents may be compared with regard to their relative standing on any given trait. For example, Respondent A may be said to be higher than Respondent B on an Achievement scale but lower than Respondent B on a Dependability scale. In contrast, an ipsative measure provides information only about the relative rank ordering of trait standings within a given respondent. For example, Respondent A may be said to be higher on Achievement than on Dependability whereas Respondent B may display the opposite ordering. What *cannot* be inferred from ipsative measures, however, is whether Respondent A is higher on either Achievement or Dependability than Respondent B. Ipsative measurement thus provides intra-individual measurement only—no inferences about respondents' relative standings on the measure are valid.

Validity of AIM for Inter-Individual Comparisons

Using Hicks's (1970) criteria, the AIM yields partially ipsative scores.[2] An important question, then, is how closely the partially ipsative scores from the AIM mirror normative scores of the same traits—in particular, the trait of Adaptability. A study by Block (1957) indicated that "there is an almost complete functional identity between the results obtained from [partially] ipsative ratings treated normatively and conventionally acquired normative rating data" (p. 52). Block's measure, however, exhibited substantial variability in the total score (1,536 units and 784 units, respectively).[3] AIM scale scores show some variability, although the score distribution is slightly less variable in the operational sample than is a similarly truncated distribution drawn

[2] This occurs because responses to AIM item stems are conditional: Once a response has been given, the response options for other stems within that item (a tetrad) are restricted. For example, a respondent cannot reply *Most Like Me* to two stems within a tetrad.

[3] Indeed, Hicks (1970) noted that Block's normative measures "actually possessed a *smaller* potential range than the 'ipsative' measures" (p. 170, emphasis added).

from the research sample. For example, assuming a normal distribution, approximately 95% of the Adaptability scores fall between 44 and 73 in the restricted research sample, and between 51 and 79 in the operational sample (cf. Chapter 3, Table 3.2).

There is evidence suggesting that the AIM is far from being a fully ipsative measure, including (a) positive scale intercorrelations, (b) relatively strong correlations with measures of will-do job performance, and (c) modest correlations with attrition that are similar in magnitude to those obtained with the Assessment of Background and Life Experiences (ABLE)—a normative measure (Mark Young, personal communication, October, 2001). Research is needed, however, to assess the extent to which partial ipsativity affects AIM scores. This research would evaluate the validity of inter-individual comparisons, which form the basis of operational accession decisions involving the AIM.

Evaluation of AIM Item Stems Using Traditional Item Statistics

A second technical issue involves the use of traditional item statistics (e.g., point-biserial or biserial correlations of item with total scale score, coefficient alpha estimates of reliability) to evaluate AIM item stems and scales. Although such statistics have been computed and documented in this report (see Chapter 7), and were used to develop AIM forms, there is some concern about whether they should be used to identify poorly functioning item stems when the data are partially ipsative.

AIM item stems appear in tetrads, and responses to these stems are partially constrained by responses to other stems in the tetrad.[4] Thus, the variability of a given stem is not independent of the other stems with which it is paired, which in turn means that both item variability and total scale score variability are conditional on other trait stems (this is part of the reason that the AIM is partially ipsative). As such, an item stem that is viewed as a poor stem in its current tetrad (e.g., it exhibits a low or even negative stem-total correlation, coefficient alpha would increase if it were removed) might not attain similar statistics if paired with (a) other stems assessing the same respective traits or (b) stems assessing different traits. Simply put, stem statistics are necessarily influenced by a context effect. For example, if one finds a low stem-total correlation, it is not possible to tell whether that low value is due to the stem being unrelated to the other stems in the scale or due to the complex relations among the stems in the tetrad. If there were no context effect, one would expect the same rank order between stem scores from the Likert and AIM scoring approaches, but that is not the case.

An alternative approach to evaluating AIM tetrads and their constituent item stems is the one used to develop the AIM alternate forms (Heggestad, Young, et al., 1999). That is, assess stem fitness (e.g., homogeneity) using data from Likert analyses. All item stems deemed to be solid indicators of the trait of interest should then be paired carefully with the stems from the other traits that the AIM measures. Presently, Trait A may be paired with Trait B in 15 tetrads but with Trait C in only 4 tetrads. Given the dependence inherent to the multidimensional forced choice (MFC) format, balancing trait appearances across tetrads is a vital step—one lacking from the current versions of the AIM. After identifying appropriate AIM stems and their pairings

[4] For a tetrad, the respondent makes two independent judgments—which stems are deemed *most like* the respondent and *least like* the respondent.

throughout the form, a decision consistency approach (see Chapter 4) could be used to determine the degree to which two or more forms yield consistent decisions for each applicant. It would also be useful to evaluate the extent to which this approach influences the degree of ipsativity in the measure.

Alternate Forms

The development and scaling of alternate forms is a difficult challenge with an instrument like the AIM. Classical item statistics have typically been an important tool for assuring the quality of new test forms, and it is unclear what might be more suitable for an instrument like the AIM. Evaluating individual item stems based on their ability to predict attrition is one possibility, but the inherent unreliability of one-item prediction is problematic, as is the fact that the item stems do not work in isolation when they are incorporated into a tetrad.

A related concern is that, to the extent that scores on the AIM are partially ipsative, the scoring process and questionable assumptions about parallel form content call into question the appropriateness of traditional scaling methods such as those used by McCloy et al. in Chapter 4. Because of these concerns, we recommend that new AIM forms be evaluated based on the degree to which they result in decisions consistent with previous forms—a procedure routinely adopted when calculating the reliability of criterion-referenced measures (e.g., Crocker & Algina, 1986). This approach examines equivalence of decisions at the cut score and could theoretically serve as a check on scaling accuracy.

Because there is not a detailed blueprint of the kind we typically see for cognitive ability and knowledge tests, it is also challenging to help ensure alternate forms are truly equivalent in content. The problem will be even more challenging if an MFS scoring process is adopted. With such a scoring system, it is unclear what items/stems are contributing to the composite score and why they do or do not make a contribution. That is, the concept of developing stems related to the six original AIM content scales in equal proportions (as was done on the current AIM forms) may lead to the development of many items/stems that are not useful. A related point about the use of MFS scoring is that it is relevant for the application at hand (in this case, predicting attrition), but the same score would likely not be as useful for predicting other criteria (e.g., job performance). Also, the ability to generate scores for each of the six content areas, which would likely be important for some applications, is lost.

Concluding Remarks

A great deal of effort has gone into the development of AIM, research support prior to its implementation as part of the GED Plus program, and the post-implementation investigations described in the present report. The post-implementation work is continuing, emphasizing in particular the pursuit of alternative scoring strategies proposed by Drasgow, Levine, and Williams. Some of these scoring strategies are likely to improve the faking resistance of AIM, which has not been quite as strong in the operational setting as expected based on evidence derived from research samples, and could address some of the concerns related to partial ipsativity raised earlier in this chapter. Additional work is also planned to further investigate information that might be used to supplement AIM in the prediction of first-term attrition (cf. Chapter 5).

Depending on developments in this ongoing work, it may also be useful to conduct research exploring various hypotheses about the differences in findings between research and operational settings that were evident in Chapters 3 and 6. Understanding these differences would facilitate the transition of new forms from development to implementation and could help forecast and/or address issues encountered in implementation. At a minimum, AIM-related analyses conducted on research data should be replicated in an operational setting before they can be considered to generalize to that setting.

The use of AIM has yielded significant strides in the broader challenge of the selection testing community to successfully implement temperament-based instruments in large-scale, high stakes assessment programs. Many lessons are being learned that will improve the Army's ability to successfully incorporate temperament assessment into screening decisions – lessons that will also generalize to other applications both within and outside the military.

POSTSCRIPT

In January 2002, the researchers in our post-implementation program participated in an all-day meeting at ARI. The meeting was held to facilitate an external review of our broader AIM program, which was conducted by four outside testing experts. This meeting provided a golden opportunity for us to review the most current data on AIM's performance under GED Plus, and explore alternative approaches for improving its operational validity.

Although several strategies for enhancing AIM's performance were explored, the Multilinear Formula Scoring (MFS) approach seemed particularly promising. Initially, the University of Illinois research team found that the MFS model developed using the research sample (see Chapter 8) did not fit the AIM data from the operational sample. Based on these findings, they recalibrated MFS scores using the operational data from GED Plus. MFS scoring significantly improved AIM's validity in the operational sample relative to the Adaptability Composite. For example, while the validity of the Adaptability Composite against 9-month attrition was only -.04, this validity rose to -.09 with MFS scoring. Importantly, the magnitude of the MFS score validity approached that of the validity originally obtained for the Adaptability Composite in the Tier 2 research sample (r = -.11).

We were also encouraged by other positive aspects of MFS scoring for AIM. For example, our analyses showed no evidence that the use of MFS scores would result in adverse impact for women or minorities. This was consistent with the findings previously reported (in Chapter 6) for the AIM Adaptability Composite. We also found evidence suggesting that MFS scoring would make AIM more difficult to fake. MFS scores had an extremely low correlation with the AIM Validity scale (r = .03) in the operational sample. This compared very favorably to the Adaptability Composite's higher correlation with the Validity scale (r = .28).

After the University of Illinois research team refined their MFS scoring procedure for the operational data, ARI spent several months developing and validating a new attrition screen using operational data from the GED Plus Program. The goal here was to develop an attrition composite with higher validity by augmenting the new MFS AIM scores with supplemental measures. Supplemental measures refer to objective measures that are routinely collected during the enlistment application process and are recorded on Army personnel databases. These types of measures were examined in Chapter 5, where they were referred to as "operational predictors."

After several iterations of model development, cross-validation, and refinement, ARI proposed a new attrition screen to replace the AIM Adaptability Composite being used under the GED Plus Program. This proposed attrition screen was briefed to LTG Dennis Cavin, Commander, U.S. Army Accessions Command in July 2002. The components of the proposed screen (which are combined using weights derived from logistic regression) include the following:

- AIM MFS score
- ASVAB Mechanical Comprehension
- ASVAB Math Knowledge

- ASVAB Verbal
- Age at application (17 – 19 vs. 20 +)
- Body Mass Index (lowest 5% or highest 5% vs. middle 90% using gender-based norms)

The validity of this new attrition composite against 9-month attrition in the operational cross-validation sample ($r = -.11$) was comparable to the validity of the Adaptability Composite in the Tier 2 research sample. Its magnitude was also nearly three times higher than that of the existing operational screen ($r = -.04$). Our findings indicate that this new attrition screen would not have adverse impact for minorities; nor would it have adverse impact for females at selection ratios of .75 or higher.

ARI's newly proposed attrition screen has been reviewed by HumRRO, RAND, and the University of Illinois research team, and we are continuing to update the operational database with new test and criterion data. We are also investigating refinements of MFS scoring that might enhance the validity of the proposed screen. This includes the use of Item Response Theory-based approaches for incorporating the supplemental measures into a composite with the MFS scores, and the adjustment of MFS scores for faking.

In sum, the efforts described in this report have contributed to the development of a new attrition screen for nongraduate applicants that performs much better than the AIM Adaptability Composite in the GED Plus Program. Further evaluation and refinement of this new measure will continue through this fiscal year (FY03).

Mark C. Young, ARI
August 2003

REFERENCES

Agresti, A. (1996). *An introduction to categorical data analysis*. New York: John Wiley & Sons.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Barnes, J. (2001a). *AIM Air Force research database codebook.* Alexandria, VA: Human Resources Research Organization.

Barnes, J. (2001b). *AIM grand research database codebook.* Alexandria, VA: Human Resources Research Organization.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44,* 1-26.

Block, J. A. (1957). A comparison between ipsative and normative ratings of personality. *Journal of Abnormal and Social Psychology, 54*, 50-54.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1997). CART *(Version 4.0).* San Diego, CA: Salford Systems.

Campbell, J.P., & Knapp, D.J. (Eds.) (2001). *Exploring the limits in personnel selection and classification*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Chernyshenko, O.S., Stark, S., Chan, K.Y., Drasgow, F., & Williams, B.A. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36,* 523 - 562.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement, 5*, 115-124.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Fort Worth: Harcourt, Brace, Jovanovich.

Drasgow, F., & Hulin, C.L. (1990). Item response theory. In M.D. Dunnette & L.M. Hough (Eds.) *Handbook of industrial and organizational psychology* (second edition, vol 1.) pp 577-636. Palo Alto, California: Consulting Psychologists Press, Inc.

Drasgow, F., Levine M. V., Tsien, S., Williams B. A., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement*, *19*, 143-165.

Dunbar, S. B., & Novick, M. R. (1988). On predicting success in training for men and women: Examples from Marine Corps clerical specialties. *Journal of Applied Psychology, 75,* 545-550.

Fisher, R. A. (1925). *Statistical methods for research workers, 1st Edition*. Edinburgh: Oliver & Boyd.

Fleiss, J.L., Cohen, J., & Everitt, B.S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72*, 323-327.

Green, D.M., & Swets, J.A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions.* Mahwah, NJ: Erlbaum.

Hambleton, R.K., & Swaminathan, H. (1983). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Hanley, J.A., & McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*, 29-36.

Heggestad, E. D., Lightfoot, M. A., & Waters, B. K. (1999). *Pre-implementation research on the Assessment of Individual Motivation (AIM) – Phase I: Item development* (FR-WATSD-99-22). Alexandria, VA: Human Resources Research Organization.

Heggestad, E.D., Young, M.C., Strickland, W.J., & Rumsey, M.G. (1999). *The Assessment of Individual Motivation: Evaluation of validity and development of alternative forms.* Technical Report (FR-WATSD-99-66). Alexandria, VA: Human Resources Research Organization.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74*, 167-184.

Hoffman, P.J. (1962). Assessment of the independent contributions of predictors. *Psychological Bulletin, 59,* 77-80.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75*, 581-595.

Houston, W. M., & Novick, M. R. (1987). Race-based differential prediction in Air Force technical training programs. *Journal of Educational Measurement, 24,* 309-320.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior, 29,* 340-362.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72-98.

Hunter, J.E., & Schmidt, F.L. (1990). *Methods of meta-analysis: Correcting for error and bias in research findings*. Newbury Park, CA: Sage.

Hunter, J. E., Schmidt, F.L., & Rauschenberger, J. (1984). Methodological and statistical issues in the study of bias in mental testing. In C. R. Reynolds & R. T. Brown (Eds.). *Perspectives on bias in mental testing*. New York: Plenum Press.

Knapp, D.J., Waters, B.K., & Heggestad, E.D. (Eds.). (2002). *Investigations related to the implementation of the Assessment of Individual Motivation (AIM)* (Study Note 2002-02). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Kolen, M.J., & Brennan, R.L. (1995). *Test equating: Methods and practices.* New York: Springer.

Langeheine, R,. & Rost, J. (1988). *Latent trait and latent class models*. New York: Plenum Press.

Levine, M.V. (2001) "Dimension in latent variable models", *Journal of Mathematical Psychology*, under review. Internet version available at URL http://www.staff.uiuc.edu/~m-levine/dimension.pdf

Levine, M.V. & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika, 53,* 161-176.

Linn, R. L., & Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement, 8*, 1-4.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: ETS Policy Information Center.

Negelkerke, N. J. D. (1991). A note on general definition of the coefficient of determination, *Biometrika, 78,* 691-692.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philisophical Transactions of the Royal Society (Series A), 231*, 289-337.

Ree, M. J., & Earles, J. A. (1991). Predicting training success: Not much more than *g*. *Personnel Psychology, 44,* 321-332.

Ree, M. J., & Earles, J. A. (1992). Intelligence is the best predictor of job performance. *Current Directions in Psychological Science, 1*, 86-89.

Rindskopf, D., & Everson, H. (1984). A comparison of models for detecting discrimination: An example from medical school admissions. *Applied Psychological Measurement, 8*, 89-106.

Saad, S., & Sackett, P.R. (2002). Investigating differential prediction by gender in employment-oriented personality measures. *Journal of Applied Psychology, 87*, 667-674.

Sackett, P. R., & Ellingson, J. E. (1997). On the effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology, 50,* 708-721.

Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist, 49*, 929-954.

Samejima, F. (1983). Some methods and approaches of estimating the operating characteristics of discrete item responses. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement*. Hillsdale, NJ: Erlbaum.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-274.

Schmidt, F. L., Pearlman, K., & Hunter, J. E. (1981). The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. *Personnel Psychology, 33*, 705-724.

Sipes, D. E, Strickland, W. J., Laurence, J. H., DiFazio, A. S., & Wetzel, E. S. (2000). *Training base attrition: Analyses and findings* (FR-00-38). Alexandria, VA: Human Resources Research Organization.

Society for Industrial and Organizational Psychology (1987). *Principles for the validation and use of personnel selection procedures*. College Park, MD: Author.

Stark, S. (2001). *MODFIT: Computer program for examining model-data fit using fit plots and chi-square statistics*. University of Illinois at Urbana-Champaign.

Stricker, L. J., Rock, D. A., & Burton, N. W. (1993). Sex differences in predictions of college grades from Scholastic Aptitude Test scores. *Journal of Educational Psychology, 85*, 710-718.

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*, 703-742.

Thissen, D. (1991). *MULTILOG user's guide (Version 6.0)*. Mooresville, IN: Scientific Software.

Trent, T., & Laurence, J. H. (Eds.). (1993). *Adaptability screening for the armed forces*. Washington, DC: Office of Assistant Secretary of Defense (Force Management and Personnel).

Valentine, L. D. (1977). *Prediction of Air Force technical training success from ASVAB and educational background (AFHRL-TR-77-18)*. Lackland Air Force Base, TX: Personnel Research Division, Air Force Human Resources Laboratory.

White, L.A., & Young, M.C. (1997, July). Prediction of attrition and will-do performance: The Assessment of Individual Motivation (AIM). Briefing presented to the Air Force Human Resources Laboratory, Brooks Air Force Base, Texas.

White, L.A., & Young, M.C. (1998, August). *Development and validation of the Assessment of Individual Motivation (AIM)*. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

White, L.A., & Young, M.C. (2001, April). *Validation of a faking-resistant measure of temperament constructs*. Paper presented at the annual meeting of the Society for Industrial/Organizational Psychology, San Diego, CA.

White, L.A., & Young, M.C., & Rumsey, M.G. (2001). Assessment of Background and Life Experiences (ABLE) implementation issues and related research. In J.P. Campbell & D.J. Knapp (Eds.), *Exploring the limits in personnel selection and classification* (pp. 526-558). Mahwah, NJ: Erlbaum.

Young, M. C., Heggestad, E. D., Rumsey, M. G., & White, L. A. (2000, August). *Army pre-implementation research findings on the Assessment of Individual Motivation (AIM).* Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

Young, M. C., & Rumsey, M. G. (1998, August). *Army pre-implementation research on the Assessment of Individual Motivation (AIM).* Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Young, M. C., & White, L. A. (1998). *Development of a measure to predict soldier attrition and motivation.* Paper presented at the 21st Army Science Conference, Norfolk, VA.

Young, M. C., White, L. A., & Heggestad, E. D. (2001, August). *Faking resistance of the Army's new non-cognitive assessment measure*. Paper presented at the annual meeting of the American Psychological Association, San Francisco, CA.

Young, M. C., White, L. A., & Oppler, S. H. (1991, October). Coaching effects on the Assessment of Background and Life Experiences (ABLE). *Proceedings of the 33rd annual conference of the Military Testing Association,* 446-451, San Antonio.

Young, M. C., White, L. A., & Oppler, S. H. (1992, October). Effects of coaching on the validity of a self-report temperament measure. *Proceedings of the 34th Annual Conference of the Military Testing Association,* 188-193, San Diego.

Zickar, M. J. (1997). Detecting untraitedness using model-based measurement. Unpublished doctoral dissertation. University of Illinois at Urbana-Champaign.

Appendix A

Moderated Regression Analysis Results

Table A-1. Moderated Regression Analysis: Army

Attrition Period	Subgroup	OLS Regression				Logistic Regression			
		Constant	Group	Adaptability	Adaptability x Group	Constant	Group	Adaptability	Adaptability x Group
3-month	Female—Male	**-0.02**	0.29	**0.00**	0.00	-1.36	1.07	-0.03	**0.00**
	Black—White	0.41	-0.08	0.00	**0.00**	**0.13**	**-0.21**	-0.03	**0.00**
	Hispanic—White	0.49	-0.16	-0.01	0.00	**0.70**	**-0.79**	-0.04	**0.00**
	Amer. In.—White	0.43	**-0.09**	-0.01	**0.00**	**0.62**	**-0.70**	-0.05	**0.01**
	Asian—White	0.55	-0.21	-0.01	**0.00**	**1.10**	**-1.18**	-0.04	**0.00**
	Other—White	0.31	**0.03**	0.00	**0.00**	**-0.28**	**0.20**	-0.03	**0.00**
6-month	Female—Male	**0.08**	0.31	**0.00**	0.00	-0.97	1.06	-0.03	**-0.01**
	Black—White	0.60	-0.13	-0.01	0.00	0.80	**-0.46**	-0.04	**0.00**
	Hispanic—White	0.67	-0.20	-0.01	0.00	1.13	-0.79	-0.04	**0.00**
	Amer. In.—White	0.60	**-0.13**	-0.01	**0.00**	**1.07**	**-0.73**	-0.05	**0.01**
	Asian—White	0.71	-0.24	-0.01	**0.00**	**1.06**	**-0.72**	**-0.03**	**-0.01**
	Other—White	0.39	**0.08**	0.00	**0.00**	**-0.08**	**0.41**	-0.03	**-0.01**
9-month	Female—Male	0.13	0.30	**0.00**	0.00	-0.79	0.99	-0.03	**0.00**
	Black—White	0.66	-0.15	-0.01	0.00	0.99	-0.56	-0.04	**0.00**
	Hispanic—White	0.71	-0.20	-0.01	0.00	1.15	-0.73	-0.04	**0.00**
	Amer. In.—White	0.58	**-0.07**	-0.01	**0.00**	**0.77**	**-0.34**	-0.04	**0.00**
	Asian—White	0.74	-0.23	-0.01	**0.00**	**1.07**	**-0.64**	-0.03	**-0.01**
	Other—White	0.41	**0.11**	**0.00**	**0.00**	**-0.12**	**0.54**	-0.02	**-0.01**

Table A-1. (continued)

Attrition Period	Subgroup	OLS Regression				Logistic Regression			
		Constant	Group	Adaptability	Adaptability x Group	Constant	Group	Adaptability	Adaptability x Group
12-month	Female—Male	0.15	0.30	0.00	0.00	-0.72	0.98	-0.03	**0.00**
	Black—White	0.67	-0.14	-0.01	0.00	0.96	**-0.49**	-0.03	**0.00**
	Hispanic—White	0.73	-0.19	-0.01	0.00	1.17	-0.70	-0.03	**0.00**
	Amer. In.—White	0.61	**-0.07**	-0.01	**0.00**	**0.80**	**-0.33**	-0.03	**0.00**
	Asian—White	0.79	-0.25	-0.01	**0.00**	**1.35**	**-0.88**	-0.03	**0.00**
	Other—White	0.42	**0.11**	**0.00**	**0.00**	**-0.11**	**0.58**	**-0.02**	**-0.01**
15-month	Female—Male	0.20	0.28	0.00	0.00	**-0.56**	0.88	-0.03	**0.00**
	Black—White	0.71	-0.15	-0.01	0.00	1.05	-0.52	-0.03	**0.00**
	Hispanic—White	0.77	-0.21	-0.01	0.00	1.35	-0.82	-0.04	**0.01**
	Amer. In.—White	0.72	**-0.15**	-0.01	**0.00**	**1.29**	**-0.75**	-0.04	**0.01**
	Asian—White	0.84	-0.28	-0.01	**0.00**	1.61	**-1.08**	-0.03	**0.00**
	Other—White	0.43	**0.13**	**0.00**	**0.00**	**-0.14**	**0.67**	**-0.02**	**-0.01**
18-month	Female—Male	0.20	0.31	0.00	0.00	-0.57	0.99	-0.03	**0.00**
	Black—White	0.73	-0.13	-0.01	**0.00**	1.04	**-0.41**	-0.03	**0.00**
	Hispanic—White	0.82	-0.22	-0.01	0.00	1.47	-0.84	-0.04	**0.01**
	Amer. In.—White	0.73	**-0.13**	-0.01	**0.00**	**1.25**	**-0.62**	-0.04	**0.01**
	Asian—White	0.88	-0.28	-0.01	**0.00**	1.63	**-1.00**	-0.03	**0.00**
	Other—White	0.42	**0.17**	**0.00**	**0.00**	**-0.19**	**0.82**	**-0.02**	**-0.01**

Table A-1. (continued)

Attrition Period	Subgroup	OLS Regression				Logistic Regression			
		Constant	Group	Adaptability	Adaptability x Group	Constant	Group	Adaptability	Adaptability x Group
3-month	Female—Male	0.24	**0.04**	-0.003	**0.00**	**0.17**	**-0.07**	-0.05	**0.01**
	Black—White	0.50	-0.13	-0.007	0.00	1.35	-0.97	-0.06	**0.01**
	Hispanic—White	0.42	**-0.06**	-0.005	**0.00**	**0.34**	**0.05**	-0.04	**-0.01**
	Other—White	0.41	-0.05	-0.005	**0.00**	**0.46**	**-0.07**	-0.04	**0.00**
6-month	Female—Male	0.30	**0.02**	-0.004	**0.00**	**0.22**	**-0.12**	-0.05	**0.01**
	Black—White	0.52	-0.12	-0.007	0.00	1.25	-0.84	-0.06	0.01
	Hispanic—White	0.46	-0.06	-0.006	**0.00**	**0.57**	**-0.17**	-0.04	**0.00**
	Other—White	0.46	-0.06	-0.006	**0.00**	**0.59**	**-0.19**	-0.04	**0.00**
9-month	Female—Male	0.32	**0.01**	-0.004	**0.00**	**0.21**	**-0.15**	-0.04	**0.01**
	Black—White	0.51	-0.10	-0.007	0.00	1.04	**-0.68**	-0.05	0.01
	Hispanic—White	0.47	**-0.06**	-0.006	**0.00**	**0.49**	**-0.14**	-0.04	**0.00**
	Other—White	0.47	-0.06	-0.006	0.00	0.59	**-0.24**	-0.04	**0.00**
12-month	Female—Male	0.33	**0.01**	-0.004	**0.00**	**0.22**	**-0.15**	-0.04	**0.01**
	Black—White	0.51	-0.10	-0.007	0.00	1.01	**-0.66**	-0.05	0.01
	Hispanic—White	0.48	**-0.06**	-0.006	**0.00**	**0.50**	**-0.15**	-0.04	**0.00**
	Other—White	0.47	-0.05	-0.005	**0.00**	**0.47**	**-0.12**	-0.04	**0.00**

Note. Bolded values are nonsignificant.

Table A-2. Moderated Regression Analysis: Air Force

		OLS Regression				Logistic Regression			
Attrition Period	Subgroup	Constant	Group	Adaptability	Adaptability x Group	Constant	Group	Adaptability	Adaptability x Group
15-month	Female—Male	0.33	**0.02**	-0.004	**0.00**	**0.13**	**-0.08**	-0.04	**0.00**
	Black—White	0.51	**-0.08**	-0.006	0.00	0.91	**-0.58**	-0.05	**0.01**
	Hispanic—White	0.48	**-0.06**	-0.006	**0.00**	**0.50**	**-0.17**	-0.04	**0.00**
	Other—White	0.47	-0.05	-0.005	**0.00**	**0.39**	**-0.06**	-0.04	**0.00**
18-month	Female—Male	0.34	**0.01**	-0.004	**0.00**	**0.16**	**-0.09**	-0.04	**0.00**
	Black—White	0.54	-0.10	-0.007	0.00	1.11	-0.74	-0.05	0.01
	Hispanic—White	0.50	**-0.06**	-0.006	**0.00**	**0.59**	**-0.22**	-0.04	**0.00**
	Other—White	0.47	**-0.04**	-0.005	**0.00**	**0.33**	**0.04**	-0.03	**0.00**
21-month	Female—Male	0.33	**0.02**	-0.004	**0.00**	**0.07**	**-0.04**	-0.04	**0.00**
	Black—White	0.54	-0.10	-0.007	0.00	1.07	-0.73	-0.05	0.01
	Hispanic—White	0.50	**-0.06**	-0.006	**0.00**	**0.52**	**-0.18**	-0.04	**0.00**
	Other——White	0.47	**-0.03**	-0.005	**0.00**	**0.26**	**0.08**	-0.03	**-0.01**
24-month	Female—Male	0.35	**0.01**	-0.004	**0.00**	**0.13**	**-0.10**	-0.04	**0.00**
	Black—White	0.56	-0.11	-0.007	0.00	1.13	-0.78	-0.05	0.02
	Hispanic—White	0.51	**-0.06**	-0.006	**0.00**	**0.52**	**-0.17**	-0.04	**0.00**
	Other—White	0.48	**-0.03**	-0.005	**0.00**	**0.25**	**0.10**	-0.03	**-0.01**

Note. Bolded values are nonsignificant.

Appendix B

AIM Item Analysis
Results for Scale C

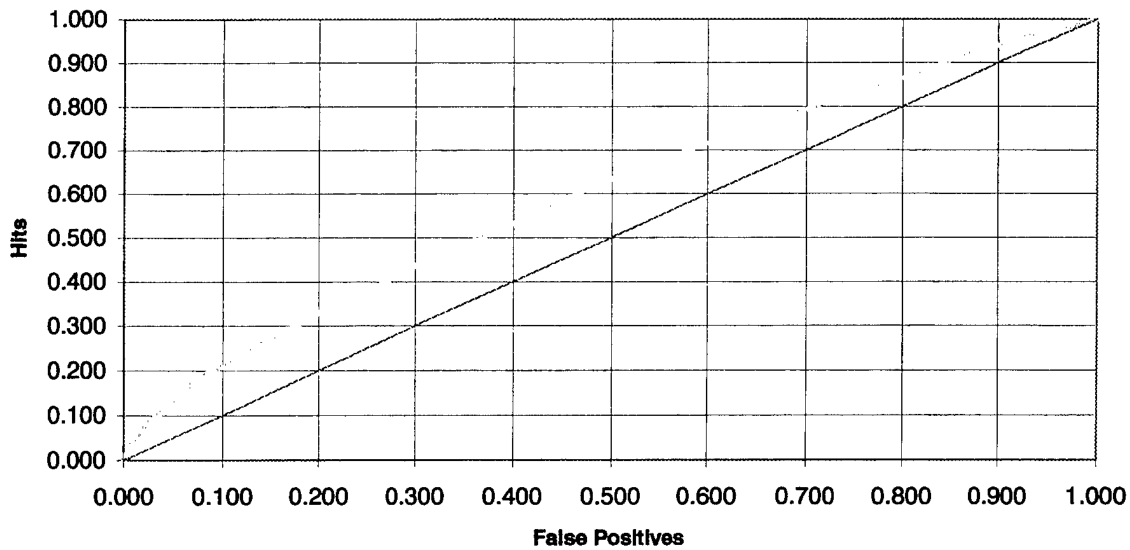
Table B.1. Classical Test Item Statistics

Scale	Alpha	Standardized Alpha	Stem (Statement)	Corrected stem-total correlations	Alpha if stem deleted
Scale C	0.682	0.683	1	0.392	0.648
			2	0.299	0.670
			3	0.431	0.638
			4	0.318	0.664
			5	0.449	0.635
			6	0.368	0.654
			7	0.315	0.665
			8	0.413	0.643

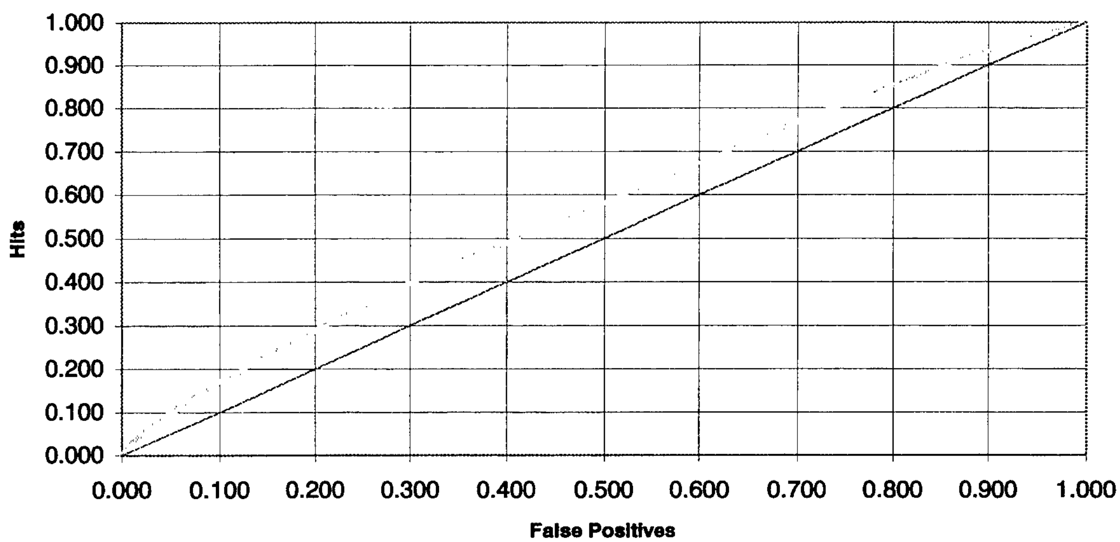
Appendix C

OAM ROC Curves for AIM Scales

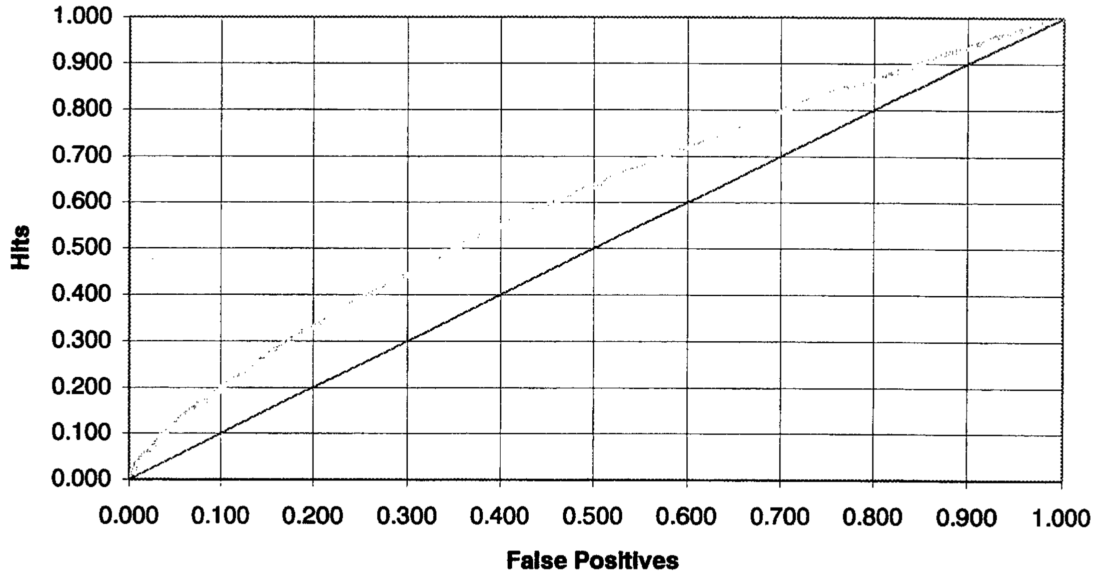
ROC for Scale A



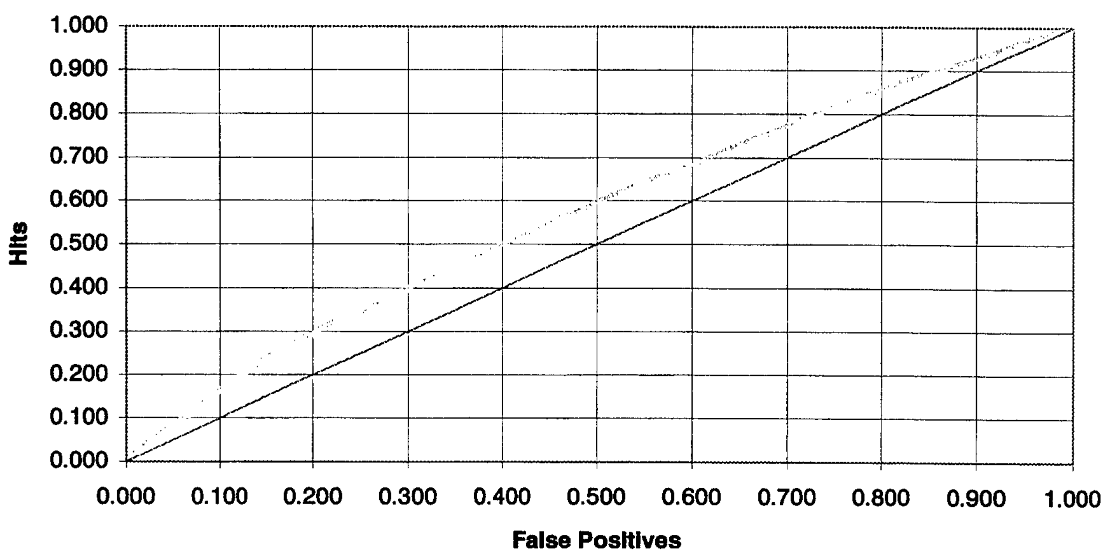
ROC for Scale B



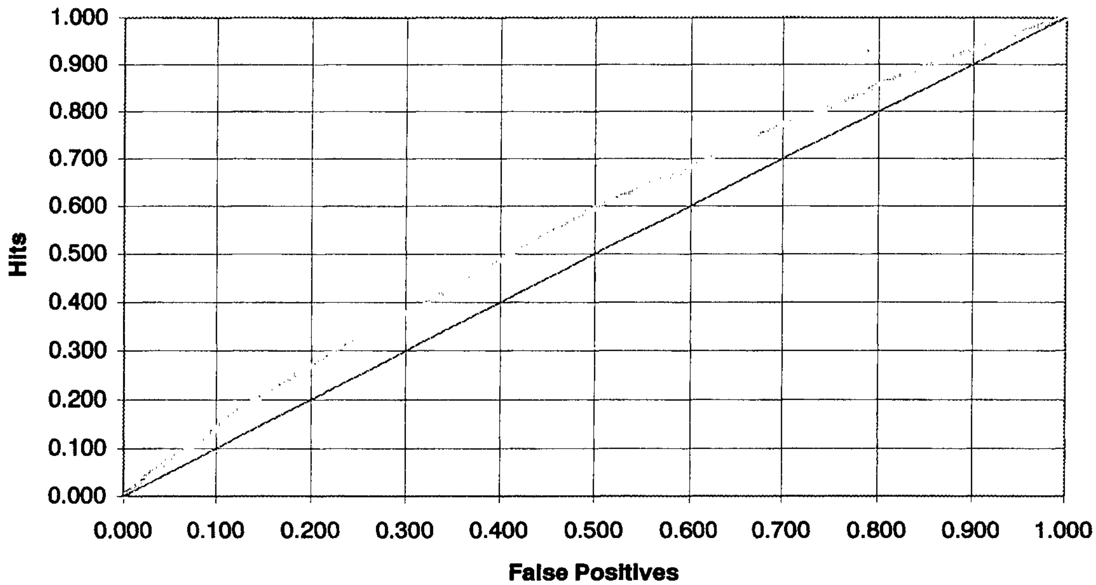
ROC for Scale C



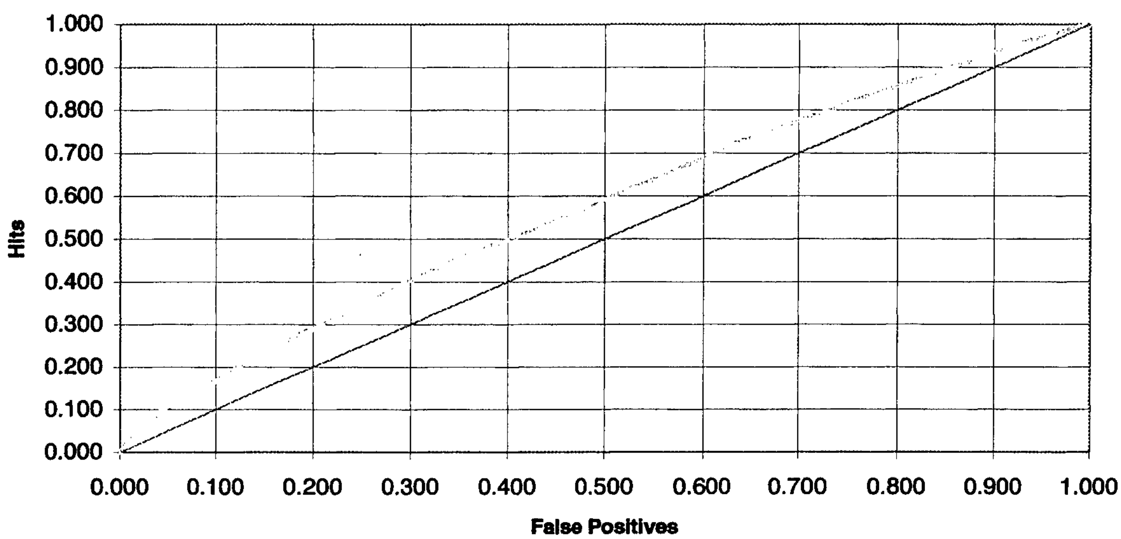
ROC for Scale D



ROC for Scale E



ROC for Scale F



ROC for Composite